UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

EXTRACÇÃO DE INFORMAÇÃO DE REGISTOS ELECTRÓNICOS DE SAÚDE

Fábio André da Cunha Guilherme

Mestrado em Engenharia Informática
Sistemas de Informação

Dissertação orientada pela:
Prof. Doutora Cátia Luísa Santana Calisto Pesquita

2016

Agradecimentos

Em primeiro lugar gostava de agradecer aos meus pais, pois sem eles nada disto seria possível e agradecer pelo apoio que me deram desde o início. Agradecer pelo exemplo de pessoas que são, pessoas que nunca desistiram de nada e que lutaram contra todas as adversidades, também nunca desistiram de lutar pelos seus sonhos, nem pelas suas vidas.

Quero também agradecer a outras pessoas que me ajudaram a realizar o meu sonho, os meus avós maternos, que apesar não estarem entre nós ajudaram-me bastante para que tivesse uma formação superior.

Queria agradecer a uma pessoa, que me ajudou mais recentemente, Cátia Sousa, por me ter emprestado o seu computador, que sem ele teria levado horas para alcançar os objectivos que tinha para cumprir.

Por último, gostava de agradecer toda a ajuda prestada, ideia da tese, à minha orientadora, Cátia Pesquita.

A todas estas pessoas, muito obrigado.

Dedico esta tese à minha família.

Resumo

Hoje em dia, a informação que circula é partilhada de uma forma eletrónica em vez, por exemplo, na forma de papel. As tradicionais fichas de paciente estão aos poucos a serem substituídas por registos eletrónicos de saúde, que facilitam a partilha de informação de médico para médico ou de clínica para clínica, dependendo da localização do paciente.

Os registos eletrónicos têm grande utilidade, por exemplo, para a prevenção de eventos adversos causados por medicamentos e no suporte à tomada de decisão. Através da análise de registos eletrónicos de saúde de um paciente e da combinação com dados clínicos é possível determinar e ajudar o médico a escolher qual o antibiótico mais indicado para o paciente, sem que seja necessário ao paciente fazer análises laboratoriais ou que o médico baseie a sua decisão apenas na sua experiência. Desta forma é possível personalizar os cuidados de saúde às necessidades e ao histórico do paciente.

Contudo as clínicas usam diferentes registos eletrónicos de saúde e cada um deles pode usar um vocabulário controlado diferente, o que aumenta a heterogeneidade da informação, o que dificulta a compreensão da informação partilhada e também dificulta a utilização dessa informação por sistemas inteligentes de apoio à decisão.

Uma forma de solucionar esse problema é criar uma estratégia que associe vários conceitos de diferentes ontologias a termos desses vocabulários controlados. Esta tese sugere uma estratégia para fazer a anotação de registos eletrónicos de saúde. Estratégia esta que recomenda quais as ontologias a usar, agrupa os termos por "paciente" e "evento", e que filtra as anotações por diferentes critérios.

Para a implementar esta estratégia foram criados 3 módulos: o EHRannotator, que faz a anotação dos registos electrónicos de saúde; o ExportConcepts, que avalia os conceitos dentro de cada ontologia recomendada e o RecommendOntologies, que faz a recomendação das ontologias. Estes módulos comunicam com uma base de dados onde estão guardados as ontologias recomendas e os registos electrónicos de saúde. Os registos electróncios de saúde são constituídos por pacientes simulados que foram retirados de uma plataforma aberta de registos electrónicos de saúde, a OpenMRS. Os módulos EHRannotator e RecommendOntologies fazem uso de dois *webservices* REST

para cumprir as suas tarefas, o Annotator e o Recommender, ambos criados pelo centro NCBO (National Center for Biomedical Ontology).

Palavras-chave: anotação semântica, ontologias, filtragem, registos eletrónicos de saúde, sistemas de suporte à decisão

Abstract

In the last decades traditional patient records have gradually been replaced by electronic health records, since they facilitate the sharing of medical information between medical practitioners and institutions.

Through the analysis of electronic health records it is possible to help doctors in different tasks, for instance performing diagnosis or prescribing a drug. These decisions aren´t based only on a doctor´s experience and knowledge but also on the patient's health history and with this information it is possible to customize care for the patient.

However the healthcare facilities routinely employ different electronic health records systems and each can use a number of different controlled vocabularies, which increases the heterogeneity of information, making it harder to share information and also to use this information for intelligent decision support systems.

One way to mitigate this problem is to increase the semantics provided by these systems by connecting the terms of controlled vocabularies to ontology concepts. This thesis suggests a strategy to perform the semantic annotation of terms in electronic health records. This strategy recommends which ontologies should be used, and filters the final annotations using different criteria to arrive at a

The implementation of this strategy was tested using an open source electronic health records system to provide insights on the usefulness and application of the proposed strategy.

To implement this strategy, 3 modules were created: the EHRannotator, which records electronic health records; The ExportConcepts, that evaluates the concepts within each recommended ontology and the RecommendOntologies, which makes the recommendation of the ontologies. These modules communicate with a database, where the recommended ontologies and electronic health records are stored. The electronic health records have simulated patients taken from an open electronic health records platform, OpenMRS. The EHRannotator and RecommendOntologies modules use two REST *webservices* to fulfill their tasks, Annotator and Recommender, both created by NCBO center (National Center for Biomedical Ontology).

Keywords: semantic annotation, ontologies, filtering, electronic health record, decision support systems

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com o passar do tempo e a evolução das tecnologias, os cuidados médicos têm vindo a criar, guardar e partilhar uma grande quantidade de dados informáticos. Estes dados podem ser informações gerais sobre pacientes, textos clínicos, medicação, diagnósticos, resultados de laboratório, imagens raio-x, etc. Actualmente a maioria desta informação é armazenada em bases de dados pelos centros de cuidados médicos, sob a forma de registos eletrónicos de saúde.

Uma das grandes áreas de investigação associada a estes dados tem a ver com o seu uso por sistemas clínicos de suporte à decisão por forma a permitir aos clínicos a tomada de decisões mais informadas e a melhorar a segurança e cuidado ao paciente e a reduzir os custos [1] [2] [3]. A análise de grandes conjuntos de registos electrónicos médicos tem também sido utilizada na investigação biomédica para facilitar a descoberta de relações entre genótipos e fenótipos, pode também ajudar os clínicos nas suas decisões, sejam elas relacionadas com tratamentos a sugerir ou que medicamento prescrever ao paciente e também na monitorização dos medicamentos após a saída para o mercado [3]. Muitas vezes, para o clínico tomar decisões tem que analisar previamente grandes quantidades de dados, tem que ter conhecimento do historial do paciente e tem de combinar toda esta informação com conhecimento clínico, o que faz com que o clínico recorra muitas das vezes à inferência. Contudo, estas decisões se forem tomadas com poucos factos podem levar a um clínico a conclusões potencialmente incorrectas. Utilizando técnicas de prospecção de dados, sobre registos electrónicos de saúde é possível melhorar as tomadas de decisão [1] [4]. Por exemplo, em situações em que o clínico tem de prescrever medicamentos, o sistema pode verificar se o medicamento prescrito pelo médico não teve qualquer efeito adverso, como alergias, no passado do paciente ou noutro paciente que tenha as mesmas características, caso não exista nenhum factor que rejeite esse medicamento, o clínico pode fazer a prescrição com maior segurança. Podem também ser utilizados para avisar previamente possíveis casos de infecção e sugerir qual o tipo de tratamento realizar,

quais os antibióticos e a dosagem a utilizar dependendo das características do paciente [3].

Já foi demonstrado que os sistemas que fazem prospeção de dados têm um desempenho melhor quando lhes são adicionados dados com relações [4], mas esta não é uma tarefa simples, e tem consistido usualmente na criação manual de regras por peritos. No entanto, as ontologias demonstram ser uma ferramenta útil para a solução deste problema. As ontologias, que são grafos de conceitos, permitem, quando estes conceitos são ligados a termos, perceber a semântica do termo. Quando os dados são ligados correctamente a conceitos (isto é, anotados) é possível perceber como é que os dados se relacionam entre si e extrair novos conhecimentos. E é por isso que as ontologias são ferramentas importantes para transformar informação não estruturada (como textos clínicos, sumários, diagnósticos, etc) ou com pouca riqueza semântica (p.ex.: classificação recorrendo a vocabulários controlados) em informação estruturada e enriquecida do ponto de vista do conhecimento que depois será utilizada pelo sistemas de prospeção de dados para extrair conhecimento.

No entanto, e se no ramo biomédico tem existido, desde há mais de uma década, um esforço para diminuir a heterogeneidade dos vocabulários e ontologias utilizados, existindo actualmente padrões utilizados em todo o mundo, o mesmo não é estritamente verdade no domínio clínico. Numa mesma unidade de saúde é usual encontrar diferentes sistemas de informação clínicos, cada um com o seu conjunto de terminologias, que muitas vezes não vão além de vocabulários controlados.

Torna-se assim importante transpor para o contexto clínico as oportunidades dadas pelo uso de ontologias mais ricas, capazes de lidar não só com a heterogeneidade dos dados, mas também facilitar a sua análise.

1.1 Objectivos

O objectivo principal desta dissertação é desenvolver novas técnicas para anotação semântica de termos utilizados em bases de dados de registos electrónicos de saúde, tendo em atenção diferentes contextos: geral, paciente e evento. Por contexto, entende-se que a anotação será orientada ao paciente, evento ou a toda a base de dados. Pretende-se também que a anotação faça uso de estratégias para selecção automática do melhor conjunto de ontologias a utilizar e também do melhor conceito a utilizar para cada anotação.

Estes objectivos levam a uma série de questões de pesquisa que orientaram este trabalho:

1. Que técnicas existentes de anotação semântica podem ser utilizadas para a anotação de registos electrónicos médicos?
2. De que forma é que estas técnicas podem ser adaptadas aos diferentes contextos?
3. Que técnicas de recomendação de ontologias para anotação podem ser adaptadas ao problema?
4. Como adaptar e melhorar as técnicas de recomendação e anotação por forma a gerar anotações completas e úteis para posterior análise?

1.2 Contribuições

Esta estratégia, inspirada nos objectivos a cumprir, permite selecionar automaticamente a ontologia ou conjunto de ontologias mais adequados para um determinado contexto, bastando apenas uma análise prévia do vocabulário controlado contido nos registos electrónicos de saúde através de um *webservice*. Esta estratégia facilita a tarefa de encontrar a ontologia ou ontologias mais adequadas, para anotar os dados contidos nos registos electrónicos de saúde.

Outra contribuição importante desta estratégia foi a criação de 3 critérios de filtragem, que resultou do facto de não existirem sistemas de anotação semântica que façam a selecção do melhor conceito para um determinado termo. Os sistemas de anotação estudados limitam-se apenas a apresentar* todos os conceitos que ligam com os termos, não existindo assim uma filtragem pelo conceito mais adequando para cada termo. Ao usarem-se estes critérios de filtragem, os termos são ligados aos melhores conceitos, o que permite anotar semânticamente dados com maior precisão.

As contribuições deste projecto são:

- Uma estratégia para descobrir a melhor ontologia ou o melhor conjunto de ontologias a utilizar para anotar termos num determinado contexto (paciente, evento) utilizando uma técnica de anotação (“texto” ou “palavra-chave”);
- Três métricas, que se baseiam na localização do conceito dentro da ontologia e no número de relações de cada conceito, e que permitem escolher o melhor conceito para cada termo;
- Um sistema que combina a estratégia para descobrir a melhor ontologia a usar para um determinado contexto e as métricas de avaliação de conceitos.

E desta fusão o sistema é capaz de anotar os termos com os melhores conceitos das melhores ontologias;

1.3 Estrutura do documento

O documento apresenta todos os desafios e resultados que foram aparecendo com o desenvolvimento desta estratégia como também toda a informação necessária para compreender a sua lógica.

No capítulo 2 serão apresentados alguns dos trabalhos desenvolvidos com registos electrónicos de saúde como também alguns conceitos relacionados com a estratégia utilizada nesta dissertação. No capítulo 3 será apresentado uma visão geral do sistema, e como é que cada um dos módulos, que constituem o sistema, trabalham entre si. No capítulo 4 será descrito como é que cada módulo foi implementado, dado ênfase ao funcionamento interno de cada um dos módulos. No capítulo 5 serão apresentados e discutidos os resultados obtidos em cada um dos módulos na anotação semântica de uma base de dados modelo de registos electrónicos médicos. O capítulo 6 apresenta as conclusões deste trabalho, bem como um delineamento de trabalho futuro.

Capítulo 2

Trabalho relacionado

Este capítulo tem como propósito apresentar alguns dos trabalhos de investigação relacionados com temas desta dissertação. Os temas vão desde a recomendação de ontologias, análise de conceitos dentro das ontologias, até à prospeção de dados dos registos electrónicos de saúde. Apesar de este último tema não ser objecto desta dissertação, mas considerando que o propósito da anotação semântica dos registos é o de permitir uma posterior aplicação de técnicas de prospecção de dados, é importante dar uma visão geral desta área.

Alguns recursos utilizados nesta dissertação, como os registos electrónicos de saúde e ontologias, serão apresentados para que o leitor perceba melhor o que são e quais as suas vantagens. Serão apresentados alguns sistemas e estratégias que usam ontologias, anotação semântica e registos eletrónicos de saúde, e que ajudam na tomada de decisão por parte do clínico.

2.1 Registos electrónicos de saúde (RES)

Os registos eletrónicos de saúde são repositórios de informação sobre um paciente num formato digital, guardados em segurança numa base de dados e acedidos por utilizadores autorizados. Foram criados com o propósito de guardar informação importante sobre o paciente e de alguma forma apoiar os médicos e outros utilizadores dos RES a melhorar a saúde dos pacientes [6]. Os registos podem guardar as mais diversas informações, como por exemplo, gráficos, administração de medicamentos, avaliação física, notas de enfermeiros, plano de cuidados, sintomas, historial médico, estilo de vida, diagnóstico, exames, tratamentos, medicação, termos de vocabulários controlados, etc… Os RES podem ser estruturados de 3 formas, orientados por tempo, onde toda a informação é ordenada cronologicamente. Na estrutura orientada ao problema, as notas vão sendo tiradas à medida que surge um problema com o paciente e

cada problema é descrito de acordo com informação subjetiva ou objetiva e avaliações ou plano. Para a estrutura orientada à fonte, a informação é agrupada tento em conta o método utilizado para extrair a informação [6].

Os RES são uma nova forma de capturar e estruturar informação sobre um paciente, transferir e partilhar informação de uma forma mais fácil e eficiente com outros especialistas. Do ponto de vista de pesquisa, mostram ser uma fonte de informação fácil de aceder e que irá permitir novas descobertas na área da saúde.

2.2 Vocabulários controlados, Ontologias e Ontologias clínicas

2.2.1 Vocabulários controlados

Os vocabulários controlados são termos, específicos de um domínio, utilizados para representar conhecimento, ou seja, são utilizados termos ou frases para catalogar e indexar informação para facilitar a pesquisa de informação. Um dos desafios é construir e controlar esses vocabulários de forma a não existirem ambiguidades nos termos, pois como consequência, podem surgir problemas que põem a saúde e a segurança dos pacientes em risco. Os vocabulários controlados clínicos podem ser de 5 tipos, linguagem natural, vocabulários, terminologias, nomenclaturas e classificações. Algumas das vantagens do uso de vocabulários controlados são ajudar na padronização de texto livre ou conteúdo estruturado, representação de observações e avaliações clinicas, codificação de testes e resultados, identificação de substâncias activas, análise de dados e suporte à decisão[1].

2.2.2 Ontologias e ontologias clínicas

Uma ontologia é uma conceptualização de um domínio, feita através do uso de conceitos e relações, que ligados entre si formam uma árvore, estruturada como uma hierarquia, formada por classes e subclasses [7]. Os componentes principais de uma ontologia são:

- Classes – é um conceito que representa um objecto para aquele domínio, cada conceito pode ter descrito por um ou mais nomes. Cada conceito é identificado por um código (IRI);
- Atributos – os atributos são as propriedades e características das classes;

[1] http://www.ctcpt.net/docs/CTC_PT_Apresentacao_20150324.pdf

• Relações – representa a forma como as diferentes classes interagem entre si (ligações is_a);

As ontologias têm qualidades como a fácil reutilização, interoperabilidade entre diferentes sistemas, a comunicação entre pessoas e/ou máquinas e ainda por exemplo, uma descrição não-ambígua de um domínio. A reutilização deve-se ao facto de que as ontologias são facilmente atualizáveis e após isso podem ser usadas de imediato. Novos conceitos e relações podem ser adicionados sempre que for preciso actualizar ou corrigir um erro. A Figura 2.1 representa uma porção da ontologia NCIT, numa representação em forma de grafo, onde os nós correspondem a classes ou conceitos, e os arcos às relações entre eles.

As ontologias clínicas vêm acrescentar algo de novo que os vocabulários controlados não têm que é a relação entre os diferentes conceitos e com isto é possível perceber como é que os conceitos se relacionam entre si e extrair novos conhecimentos.

No projecto desta dissertação são utilizadas várias ontologias e vocabulários clínicos como NCIT[2], que é um vocabulário para cuidados clínicos, investigação básica e translacional, e actividades administrativas; a ontologia LOINC[3], que é uma terminologia comum para laboratório e observações clínicas; HL7[4], é um vocabulário com termos relacionados com o domínio das tecnologias da informação de saúde; RCD[5], ontologia com termos e códigos clínicos; OGG-MM[6], é constituída por genes e genomas de organismos biológicos. Todas elas estão relacionadas com a área da saúde.

[2] http://bioportal.bioontology.org/ontologies/NCIT
[3] http://bioportal.bioontology.org/ontologies/LOINC
[4] http://bioportal.bioontology.org/ontologies/HL7
[5] http://bioportal.bioontology.org/ontologies/RCD
[6] http://bioportal.bioontology.org/ontologies/OGG-MM

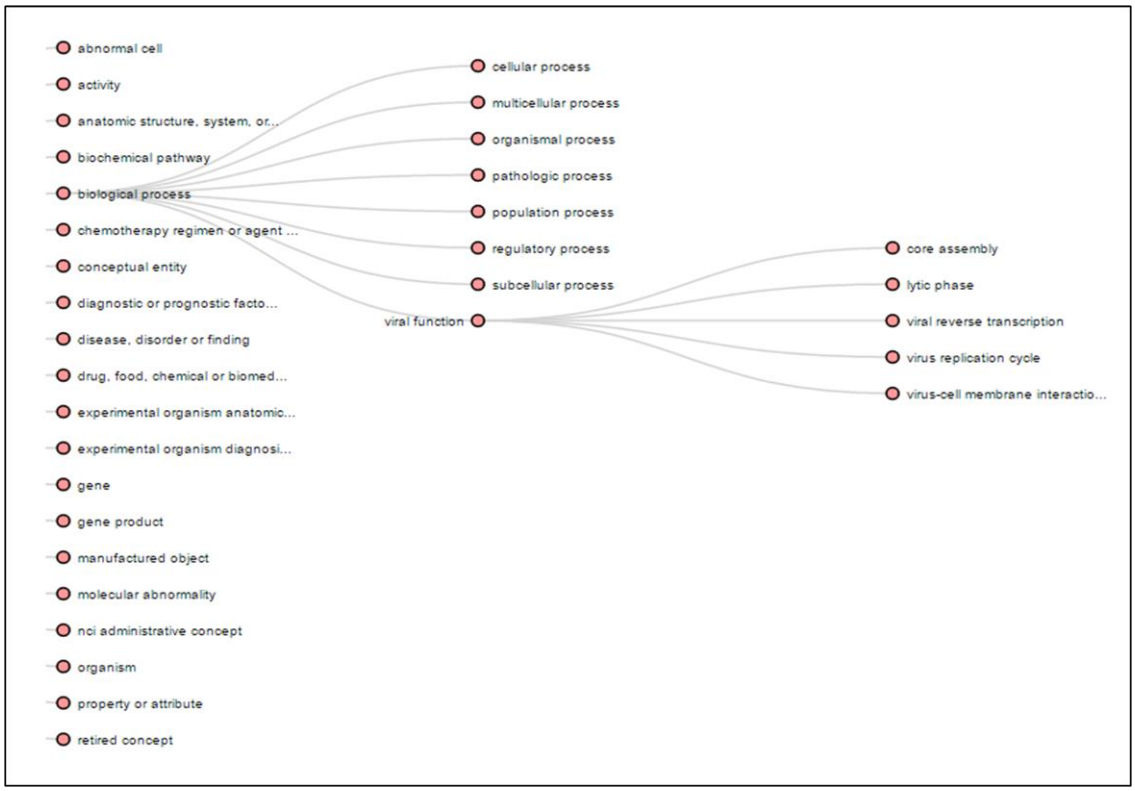

Figura 2.1- Amostra do Grafo de conceitos da ontologia NCIT

2.3 Anotação semântica

O processo de anotação semântica consiste em, simplesmente, associar um objecto a uma classe de uma ontologia. Este objecto pode ser um termo textual, um documento, ou qualquer outra entidade. Com isto o termo ganha um significado e relações com outros conceitos [8]. Na Figura 2.2 podemos ver que a palavra Bulgária, que se encontra a laranja na frase, refere-se a um país porque foi ligada à classe Bulgária que pertence a "país" e está relacionada com "localização".

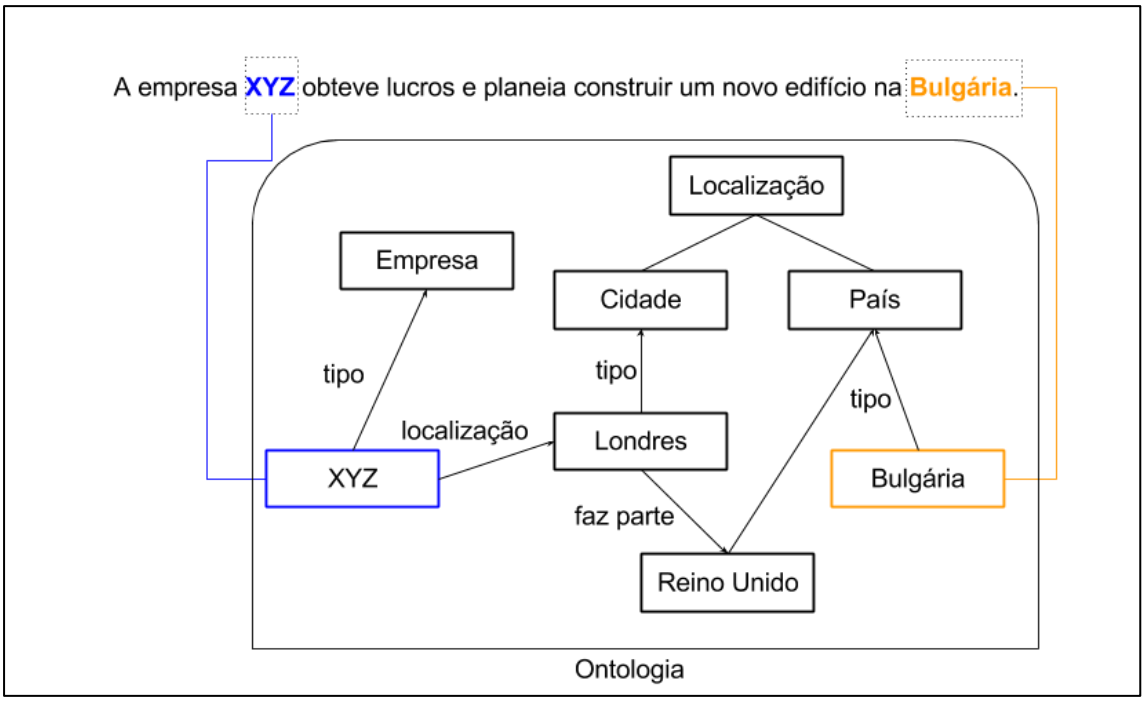

Figura 2.2- Anotação semântica (baseado numa imagem http://www. ontotext.com)

A interoperabilidade é uma grande vantagem dada pelo uso de anotação semântica, uma vez que a comunicação e interação entre diferentes sistemas é mais fácil porque ambos usam a mesma base, neste caso, usam a mesma ontologia, desta forma não é necessário nenhum passo intermédio para que a comunicação se realize com sucesso.

No caso da recolha de informação, feita por motores de busca, o problema que algumas vezes surge é a existência de resultados ambíguos, com o uso de anotação semântica esse problema estaria resolvido. A título de exemplo, vejamos o caso em que se faz uma pesquisa por "niger", o resultado serão todos os documentos relacionados com a palavra "niger", criando um problema. Existe um rio e um país com o mesmo nome e por isso os resultados apresentados serão relacionados com os dois, criando assim uma ambiguidade. Usando um motor que fizesse uma procura por anotações semânticas relacionadas com o rio "niger", os resultados apresentados seriam apenas relacionados com o rio, solucionado assim a ambiguidade [8]. As subsecções seguintes apresentam duas ferramentas, o Annotator que faz anotação semântica de termos clínicos, e o cTAKES que faz a anotação semântica de textos clínicos.

2.3.1 Annotator

Outra das aplicações disponíveis no **NCBO** é a aplicação web Annotator [9], que também está disponível na forma de webservice. Esta aplicação serve para anotar texto biomédico com conceitos provenientes de mais de 529 ontologias biomédicas vindas da UMLS e do NCBO BioPortal.

O funcionamento da aplicação é muito simples: os utilizadores inserem um texto para ser anotado. Este texto é processado com uma ferramenta de reconhecimento de conceitos, a mgrep e um dicionário. O dicionário é construído a partir dos nomes dos conceitos dentro das ontologias e dos sinónimos dos conceitos. O mgrep vai combinar os conceitos e sinónimos com termos dentro do texto que foi dado como *input* e criar uma lista de anotações directas. Na fase final, após terem sido criadas as anotações directas, são usadas 3 componentes, a componente de expansão semântica, fecho transitivo (is_a). A expansão das anotações é feita juntando os pais dos conceitos às anotações. Na componente de distância semântica, são usadas medidas de distância semântica entre os conceitos para acrescentar novas anotações ao conjunto de anotações directas, e a componente de mapeamento de ontologias é utilizada para acrescentar conceitos mapeados noutras ontologias. O resultado final é uma lista de anotações, constituída por anotações directas, e caso tenha sido pedido pelo utilizador, a lista terá também anotações expandidas.

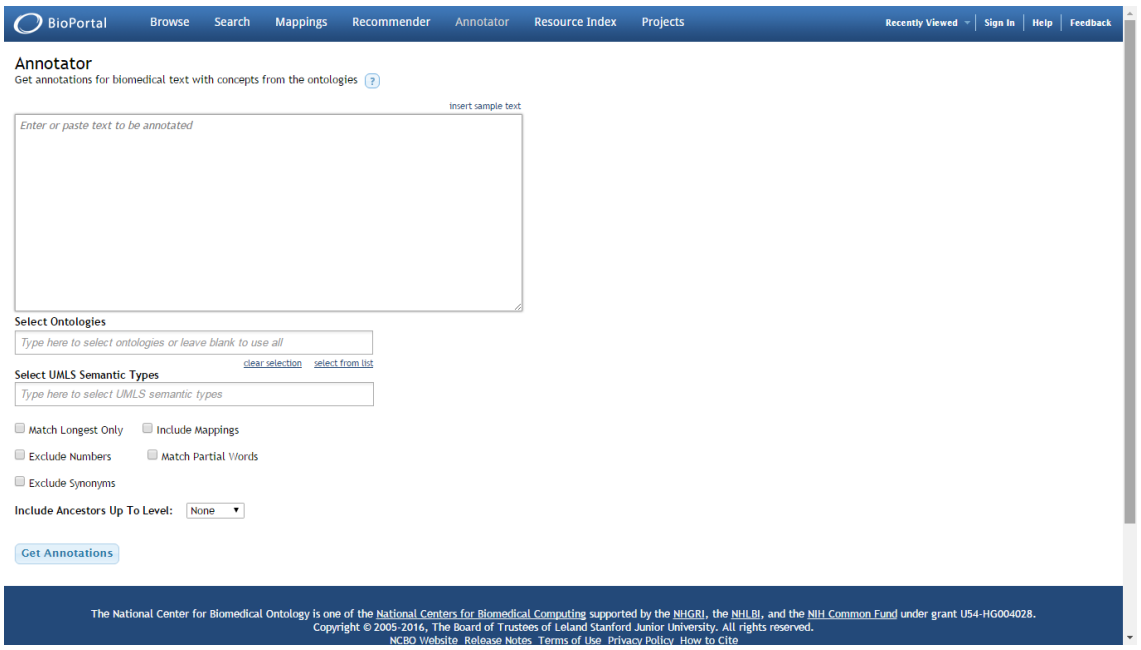

Figura 2.3- Interface do Annotator

CLASS filter	ONTOLOGY filter	TYPE filter	CONTEXT	MATCHED CLASS filter	MATCHED ONTOLOGY filter
animal cell	BioModels Ontology	direct	**animal cell**	animal cell	BioModels Ontology
animal cell	Ontology for Newborn Screening Follow-up and Translational Research	direct	**animal cell**	animal cell	Ontology for Newborn Screening Follow-up and Translational Research
cellule animale	Ontology of Nuclear Toxicity	direct	**animal cell**	cellule animale	Ontology of Nuclear Toxicity
animal cell	Cell Line Ontology	direct	**animal cell**	animal cell	Cell Line Ontology
animal cell	Cell Ontology	direct	**animal cell**	animal cell	Cell Ontology
animal cell	Brucellosis Ontology	direct	**animal cell**	animal cell	Brucellosis Ontology
animal cell	Ontology of Drug Neuropathy Adverse Events	direct	**animal cell**	animal cell	Ontology of Drug Neuropathy Adverse Events
animal cell	Uber Anatomy Ontology	direct	**animal cell**	animal cell	Uber Anatomy Ontology
animal cell	Beta Cell Genomics Ontology	direct	**animal cell**	animal cell	Beta Cell Genomics Ontology
animal cell	SMART Protocols	direct	**animal cell**	animal cell	SMART Protocols
animal cell	Porifera Ontology	direct	**animal cell**	animal cell	Porifera Ontology
animal cell	go-plus	direct	**animal cell**	animal cell	go-plus
animal cell	Neglected Tropical Disease Ontology	direct	**animal cell**	animal cell	Neglected Tropical Disease Ontology
animal cell	DebugIT Core Ontology	direct	**animal cell**	animal cell	DebugIT Core Ontology
animal cell	BioTop Ontology	direct	**animal cell**	animal cell	BioTop Ontology
multicellular organism	Cell Line Ontology	direct	**animal** cell	multicellular organism	Cell Line Ontology
multicellular organism	Ontology of Vaccine Adverse Events	direct	**animal** cell	multicellular organism	Ontology of Vaccine Adverse Events
multicellular organism	Cell Ontology	direct	**animal** cell	multicellular organism	Cell Ontology

Figura 2.4- Lista de anotações criadas pelo Annotator

Como é possível ver pela Figura 2.3 o utilizador pode configurar o processo de anotação, pode escolher as ontologias a usar, a forma como é que os conceitos são combinados com os termos, expandir as anotações usando *mapping* ou os conceitos ancestrais dos conceitos, usar ou não sinónimos. A Figura 2.4 representa uma lista de anotações criada pelo Annotator.

2.3.2 cTAKES, sistema de extracção de informação semântica

É um sistema que, através do processamento de linguagem natural, processa e extrai conteúdo semântico dos registos electrónicos de saúde. O objectivo dos criadores do sistema é dar suporte à investigação clínica, que necessita de um sistema robusto, escalável, capaz de satisfazer as necessidades na área da investigação clínica [10].

É composto por vários módulos ligados entre si e que combinam o uso de regras com aprendizagem automática para extrair informação das narrativas clínicas. Cada componente é executada em série para processar a narrativa clínica, onde cada uma contribui para a amostra semântica final. As diferentes componentes são as seguintes:

- Detector de limites de frase;
- *Tokenizer*;
- Normalizador;
- *Tagger* de discurso;
- *Shallow parser*;
- Anotador;

Cada componente tem a sua função: o detector de limites verifica onde terminam e começam as frases, procura por pontos de final ou pontos de interrogação; O *Tokenizer*, separa os termos dentro dessas frases pelos espaços ou pontuações; O normalizador, cria um invólucro à volta de cada termo, o que permite representar cada

termo do *input*; O shallow parser e o tagger são componentes utilizados no módulo de processamento de linguagem natural; O anotador é responsável por fazer a anotação semântica, que consiste em fazer o mapeamento de conceitos dentro de ontologias (como o SNOMED-CT) a termos dentro do *input*.

2.4 Avaliação e recomendação de ontologias

Nos tempos de hoje, a engenharia de ontologias está a ganhar mais popularidade e com isso cada vez há mais pessoas a criarem as suas ontologias para as usarem nos seus projectos. São criados plataformas na *internet*, como o BioPortal, onde são guardadas ontologias para que os utilizadores as possam analisar. Também fornecem outras ferramentas relacionadas com ontologias e anotação semântica, impulsionando assim o uso e a criação de ontologias, umas mais especializadas que outras num determinado domínio, mais ricas ou menos ricas em conceitos ou em relações.

Dado o esforço e custo envolvidos no desenvolvimento de uma ontologia, e também as vantagens da reutilização de uma já existente, é comum que a reutilização seja a opção tomada no desenvolvimento de sistemas com base semântica.

Optando por usar uma ontologia que já exista, outros problemas são levantados, como o de escolher a melhor ontologia. Importa aqui referir que podem existir várias ontologias com domínios relevantes para o problema, e que diferentes ontologias têm diferentes características a vários níveis, que podem ter um impacto no seu uso aplicado. Os trabalhos que serão em seguida apresentados tentam solucionar estes problemas.

2.4.1 OntoKhoj

OntoKhoj [11] é um portal semântico na *internet* que ajuda os utilizadores a classificar, avaliar e procurar ontologias. A ferramenta de classificação vê a ontologia como um texto simples que contém conceitos e relações. O modelo de classificação foi treinado usando dados do DMOZ (que é um diretório-humano onde as pessoas podem catalogar páginas da internet), ajudando assim a determinar os tópicos das ontologias.

O avaliador de ontologias baseia-se em *hyperlinks* (que referenciam a ontologia), na identidade e o lugar de quem a refere, número de citações, a distância entre conceitos para classificar a ontologia.

Para pesquisar, o **Ontokhoj** usa duas interfaces. A interface de interrogações orientada ao contexto, feita para ser usada por humanos, tem 3 dicionários de entrada, sentidos, sinónimos e hiperónimos, criando assim resultados de pesquisa mais precisos.

A segunda interface, interface máquina, foi feita para ser usada por sistemas, agentes que acedem e procuram no diretório de ontologias classificadas.

2.4.2 AKTiveRank

Este sistema, o **AKTiveRank** [12], avalia uma ontologia baseando-se na análise da sua estrutura. O utilizador envia uma interrogação HTTP para o programa, esta contém termos que irão ser comparados com conceitos (classes) das ontologias. O programa assim que recebe a interrogação pergunta ao SWOOGLE por ontologias que contenham classes com esses termos e o resultado será uma lista de ontologias com os respetivos URI. No próximo passo o programa vai verificar se essas ontologias existem na base de dados Jena, e se não existiram, são transferidas para a base de dados. A biblioteca Jena é utilizada para fazer a análise à estrutura das ontologias. O programa faz uso também de um *servlet*, o jung, que permite analisar e visualizar a estrutura da ontologia como um grafo.

Finalmente, o programa analisa cada ontologia candidata para determinar qual é a mais relevante tendo em conta a interrogação utilizada e faz uso das seguintes métricas que avaliam as ontologias:

- ***Class Match Measure*** **(cmm**): mede a cobertura que uma ontologia faz sobre os termos da *querie*. O programa tenta encontrar nomes de classes nas ontologias que sejam totalmente iguais ou parcialmente iguais aos termos da pesquisa;

- **Medida de Densidade:** esta medida avalia como é que a ontologia especifica o conceito (número de subclasses), número de atributos associados ao conceito, número de sinónimos. Avalia como é que o conhecimento sobre esse conceito está representado na ontologia e para isso analisa o número de relações, subclasses, superclasses e sinónimos;

- **Medida de Semelhança Semântica:** Esta medida tira partido do facto de que as ontologias podem ser analisadas como grafos semânticos e aplica medidas de semelhança semântica para analisar os conceitos no grafo. Esta medida calcula a proximidade dos conceitos, que fazem correspondência com os termos da pesquisa. **Ideia:** ontologias com conceitos muito afastados uns dos outros não devem representar muito bem o conhecimento de uma forma coerente e compacta;

- **Medida *Betweeness*:** Este algoritmo calcula o número de caminhos curtos que passam por cada nó do grafo. Nós que apareçam em muitos caminhos curtos têm um grande valor de *betweeness*. **Ideia:** uma classe que tenha um grande valor de *betweeness*, então é uma classe muito central na ontologia. Logo ontologias que tenham classes centrais associadas a termos de pesquisa, irão ter valores altos de *betweeness*;

2.4.3 NCBO Recommender

O **NCBO Recommender** [13] [14] é uma aplicação web e um *webservice* disponível no BioPortal. O seu propósito é recomendar a melhor ontologia ou ontologias biomédicas para um texto dado pelo utilizador com conteúdo biomédico. Estas ontologias, segundo o Recommender, serão as melhores ontologias a utilizar para anotar e representar o *input*.

O utilizador dá como *input* ao programa um texto ou um conjunto de palavras-chaves e o programa irá analisar estas palavras juntamente com o Annotator, outro programa fornecido pelo NCBO, e devolver uma lista de ontologias. A diferença de analisar os termos como "palavras-chaves" ou "texto", está na forma como o Recommender os analisa. No caso de "palavras-chave" e como estas são separadas por vírgulas, o Recommender tenta combinar conceitos com todo o termo contido entre as vírgulas e não apenas com palavras soltas entre as vírgulas, por isso é que a anotação com palavras-chaves é mais difícil e apresenta menos anotações (Annotator). Se a análise for feita por "texto", o Recommender vai tentar combinar conceitos com qualquer termo dentro das vírgulas ou com qualquer termo dentro do texto enviado ao Recommender.

Neste processo, o Annotator é usado para criar anotações sobre o texto inserido; depois são usadas métricas que avaliam as ontologias. As métricas são as seguintes:

- **Métrica de cobertura:** avalia a capacidade de uma ontologia ou ontologias em gerar anotações que cobram os termos do texto e ao mesmo tempo reflecte quantas anotações é que foram usadas do conjunto criado através da utilização de todas as ontologias. Tem em conta a qualidade das anotações geradas;

- **Métrica de aceitação:** serve para ver o quão conhecida e confiada uma ontologia é, baseia-se no número de visitas à página da ontologia e na ausência ou presença dos umls da ontologia;

- **Métrica de detalhe:** a nota do detalhe do conhecimento é calculada usando o número de definições, sinónimos, propriedades dos conceitos existentes na ontologia que cobre os termos no texto inserido;

- **Métrica de especialização:** esta métrica calcula o quão espacializada é uma ontologia para o texto inserido, para isso são usados os números e tipos de anotações encontradas pelo Annotator e a posição dentro da ontologia de cada conceito utilizado para anotar, com isto é possível saber se uma dada ontologia especifica da melhor forma o conhecimento para o texto inserido.

No final é criada uma avaliação final para cada ontologia, que é calculada somando todas as 4 métricas e multiplicando cada uma delas por um peso diferente. No caso do programa, por defeito, o peso da cobertura é de 0.55, da aceitação 0.15, do detalhe 0.15 e o peso da especialização é de 0.15. Com isto é formada uma lista com todas as ontologias recomendadas e ordenadas pela avaliação final.

O processo de recomendação pode ser configurado de diferentes maneiras: o utilizador pode escolher como é que o texto é analisado, como “palavras-chaves” ou “texto”, se a lista de recomendações é constituída por ontologias individuais ou conjuntos de ontologias, pode definir os pesos a atribuir a cada métrica e escolher as ontologias usadas no processo. Na Figura 2.5 pode-se ver a interface do Recommender e na Figura 2.6 pode-se ver as ontologias recomendadas.

Figura 2.5- Interface do Recommender

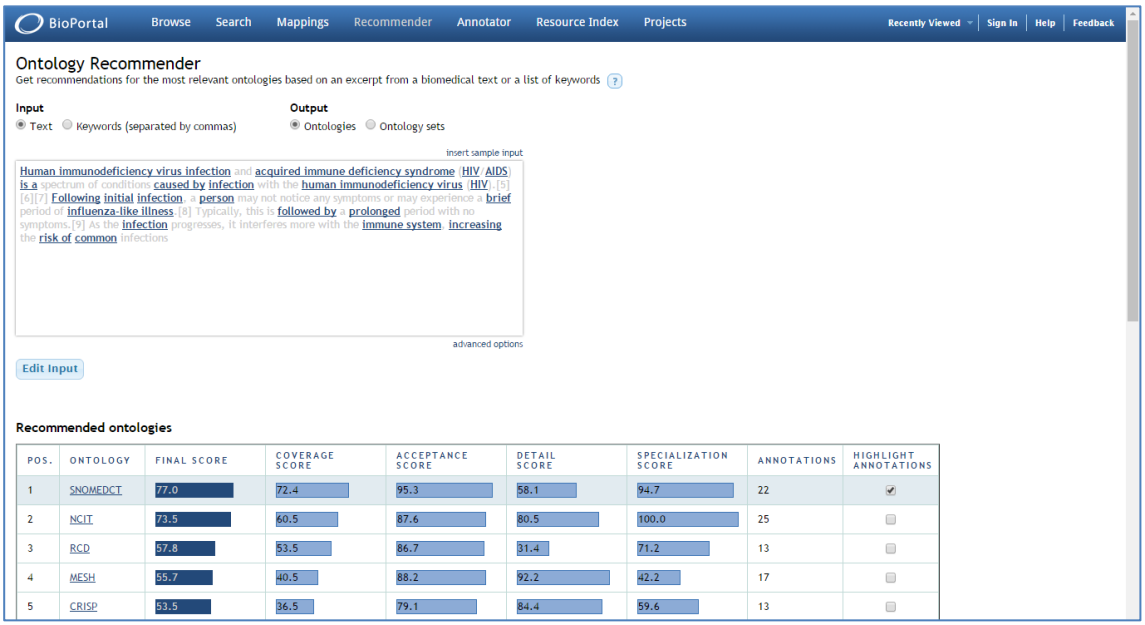

Figura 2.6- Resultado da execução do Recommender

No projecto, para avaliar uma ontologia ou ontologias, foi utilizado o Recommender. As métricas utilizadas pelo Recommender para avaliar ontologias são diferentes das que são usadas pelo AKTiveRank. Por exemplo a métrica de cobertura, é um pouco semelhante ao do AKTiveRank, mas tem em conta a qualidade das anotações, o que não é feito pelo AKTiveRank. As métricas do Recommender são suficientes e

mais indicadas para avaliar uma ontologia do que as que são utilizadas pelo AKTiveRank. As métricas do Recommender mostram se a informação da ontologia é adequada ou não e se a informação está bem detalhada e especificada.

2.5 Avaliar conceitos dentro de uma ontologia

Várias questões foram levantadas durante o desenvolvimento deste trabalho: "Como descobrir a melhor forma para avaliar um conceito dentro de uma ontologia?", "Como comparar conceitos de diferentes ontologias?". Os algoritmos que se seguem serviram de inspiração para desenvolver métricas que ajudassem a resolver esses problemas.

2.5.1 DWrank

DWrank [15] é um algoritmo bidirecional de duas fases para avaliar conceitos numa ontologia. O programa utiliza duas medidas para avaliar os conceitos, Hubscore, que avalia a centralidade do conceito dentro da ontologia e o AuthorityScore, que avalia a autoridade da ontologia. O programa usa um algoritmo de aprendizagem por classificação, que é utilizado para distribuir o peso sobre os dois tipos de avaliação anteriores. O algoritmo é de duas fases porque tem uma parte *online*, utilizada pelo utilizador para fazer as suas *queries*, e uma parte *offline*, onde o processo de aprendizagem acontece, nesta fase o programa aprende a distribuir os pesos sobre as pontuações de centralidade e autoridade.

Muitos dos sistemas de classificação que estão agora disponíveis apenas consideram a popularidade da ontologia e para isso usam algoritmos *PageRank* (como o do Google), o que penaliza ontologias que estão a emergir e que possivelmente, podem estar mais bem definidas. Penalizando, também, os conceitos dentro destas ontologias menos populares.

Neste artigo os autores apresentam uma plataforma que usa um número de técnicas para avaliar e ordenar cada conceito dentro de uma ontologia baseando-se na sua localização dentro da ontologia e na autoridade da ontologia onde se encontram.

Primeiro o programa calcula o valor de centralidade de um conceito dentro da ontologia, para isso faz uso das ligações desse conceito a outros conceitos. Depois o valor de autoridade do conceito é calculado. Para obter o valor de autoridade são usadas

as ligações entre as ontologias. Este último valor baseia-se na ideia de que uma ontologia tem uma grande autoridade caso seja usada por outras ontologias. Finalmente uma lista de conceitos é apresentada ao utilizador com os conceitos ordenados pela pontuação calculada pelo DWrank.

Os conceitos são avaliados tendo em conta a sua centralidade e a sua autoridade. Um conceito é mais central numa ontologia se tiver muitas relações a começarem dele (conectividade). Um conceito também é mais central se tiver uma relação a começar nele e ir para um conceito central (vizinhança). Os autores adoptam um algoritmo *ReversePageRank*, onde um conceito é mais "popular" se tiver caminhos (referências) para outros conceitos, do que um conceito com muitos caminhos (referências) para ele.

A autoridade de um conceito depende também da autoridade da ontologia onde este conceito está. Uma ontologia é mais autoritária se existirem mais ligações de outras ontologias a acabarem nela, ou seja, esta ontologia é referenciada por outras ontologias (re-utilização). Uma ontologia é autoritária se existir uma ligação a começar numa ontologia autoritária e a acabar nela (vizinhança).

2.5.2 Algoritmo que identifica os conceitos chaves numa ontologia

Peroni et al [16] descreve uma forma de identificar os conceitos chaves dentro de uma ontologia e assim sumarizar o tema de que a ontologia trata. Esta fórmula é inspirada no problema bastante comum e que afecta quem tem a necessidade de usar ontologias para os seus projectos e que consiste em saber o domínio de uma determinada ontologia. Sabendo mais informação sobre uma ontologia, a tarefa de descobrir se aquela ontologia é adequada para o seu projecto fica mais fácil.

Os autores defendem que esta fórmula será útil para outras áreas como selecção de ontologias, classificação automática, modularização de ontologias, e avaliação de ontologias.

Os conceitos chave, ou categorias naturais, serão as palavras onde todas as outras palavras irão pertencer, ou a maior parte das palavras. Os conceitos chave são termos que não são muito abstractos, nem muito específicos.

Os autores chegaram as várias métricas que ajudam a avaliar os conceitos dentro de uma ontologia e assim identificar quais são os conceitos-chave. As métricas são as seguintes:

- Simplicidade do nome: um conceito tem um nome simples se contiver poucos termos, quantos mais termos tiver menor será a sua nota.

- Nível básico: esta métrica avalia o conceito em relação à sua centralidade. Para calcular o valor central de um conceito é necessário saber quais são os caminhos que contêm esse conceito e em quais desses caminhos é que o conceito está no meio.

- Densidade global do conceito: esta métrica avalia o conceito em relação à sua densidade, por outras palavras, avalia como é que a ontologia descreve esse conceito na globalidade. Para isso usa o número de relações desse conceito e o número máximo de relação dentro da ontologia.

- Densidade local: é semelhante à densidade global, mas neste caso limita-se aos conceitos que rodeiam o conceito que se está a estudar, por isso em vez de se usar o número máximo de relações da ontologia, usa-se o número máximo de relações na vizinhança do conceito.

- Cobertura: Os conceitos escolhidos, ou o conceito tem que ter uma boa cobertura sobre a ontologia. Para calcular este valor basta usar o número de descendentes do conceito e o número total de conceitos na ontologia.

2.6 Prospecção de dados e RES

Apesar do presente trabalho não incidir directamente sobre prospecção de dados em RES, importa compreender o seu contexto por forma a ter uma melhor percepção dos desafios de anotação para uma prospecção posterior.

2.6.1 Monitorização de possíveis eventos adversos provocados por medicamentos

Neste artigo [17] os autores explicam como é que através de técnicas de prospecção de dados a um texto que representa vários sumários de diferentes registos electrónicos de saúde, conseguem criar uma lista de pontos de eventos adversos criados por

medicamentos (ADE) úteis para monitorizar a saúde do paciente. Por outras palavras, os autores combinam tecnologias de prospeção de dados com registos electrónicos de saúde e com anotação semântica para extrair informação. O objectivo final será usar estes padrões/informação para facilitar as tarefas de monitorizar e prever possíveis eventos adversos provocados por medicamentos nas pessoas.

Tem como vantagens reduzir os custos de pesquisa e de desenvolvimento de novos medicamentos aos laboratórios de medicamentos e também, reduzir o número de casos ADE nos pacientes e por sua vez melhor a qualidade de saúde.

Os autores defendem que as tecnologias semânticas são deveras importantes porque permitem transformar os dados em dados semânticos, ou seja, dados que têm significado e sendo dados, podem ser processados e lidos por máquinas de maneira a criarem informação, descobrirem novas ligações entre os dados originais e assim melhorar a descoberta de conhecimento.

Os dois algoritmos propostos são os seguintes:

• ***Ontology-based k-itemset enrichment*** **(OKE)** – neste algoritmo, além dos k-itemsets gerados através de métodos de prospeção, o algoritmo descobre outros k-itemsets a partir dos dados originais através da exploração e integração de outras relações da ontologia. O algoritmo não irá ter em conta apenas as relações is_a, irá dar ênfase a relações associadas a ADE´s, como: *evidendeOfADE*, *mechanismOfMolecularFunction*, e *causeOfADE*.

• ***Semantic hypergraph-based k-itemset generation*** **(SHKG)** – Este algoritmo permite descobrir novos nós através da análise de hipergrafos.

2.6.2 Lista de eventos a vigiar

Visto que fazer prospeção de dados sobre os RES tem vindo a ser mais popular e uma ferramenta importante na vigilância dos medicamentos após a sua venda, os autores deste artigo decidiram criar uma lista ordenada com eventos adversos a medicamentos [18].

Dada a grande quantidade de RES e à sua disponibilidade, é quase possível obter sinais de eventos adversos a medicamentos em tempo real. A importância de vigiar os fármacos para possíveis eventos adversos após a sua venda é muito grande, pois quanto mais cedo se intervir, menor será o número de pessoas a usarem o medicamento, e assim, menor será o número de acidentes relacionados com esse medicamento.

A ideia base foi criar esta lista para ajudar a descobrir quais é que seriam os eventos mais importantes e para estes, identificar sinais que ajudem a identifica-los.

Desta forma a tarefa de vigiar e prever estes eventos usando prospeção de dados sobre os RES será mais eficiente.

O método de criação da lista é constituído por várias tarefas. Primeira tarefa, identificar eventos importantes. Nesta tarefa a equipa teve que identificar todos os eventos adversos a medicamentos a partir de livros, artigos publicados, e informações em páginas de agências reguladoras de medicamentos na internet. Segunda tarefa, criar um conjunto de critérios para avaliar os eventos. Nesta tarefa os autores criaram o critério que mede a frequência em que há um evento que retira o medicamento do mercado (*withdrawal*), criaram um critério que mede a frequência com que o evento põe o medicamento na caixa preta de aviso e leva uma pessoa para as urgências, outro critério que mede a probabilidade do evento estar relacionado com um medicamento e criaram um critério que mede a probabilidade do evento causar a morte de uma pessoa. Todos os critérios foram avaliados de 0 a 3. A terceira tarefa consiste em avaliar cada um dos eventos. A quarta e última tarefa consiste em ordenar a lista dos eventos, sendo que os eventos em primeiro ligar são os eventos que têm prioridade na vigilância.

No topo da lista estão os seguintes eventos: erupções cutâneas bolhosas, insuficiência renal aguda, choque anafilático, enfarte agudo do miocárdio e rabdomiólise. Logo, são para estes eventos que se tem de concentrar toda a atenção e descobrir sinais que ajudem a monitorizá-los.

2.6.3 Consórcio SHARPn

Neste artigo [19] é apresentada uma plataforma que permite normalizar tanto os dados estruturados, como os não estruturados dos RES e transformá-los num modelo conceptual que facilite a extração de fenótipos.

O programa SHARP, desenvolvido pelo HIT (The Office of the National Coordinator for Health Information Technology), tem 4 áreas diferentes, Sharps, responsável pela segurança, SHARPc, responsável pelo suporte cognitivo centrado no paciente, SMART, focada nas aplicações de saúde e desenho de redes, por último a SHARPn, é responsável pelo uso secundário dos RES.

Os objetivos desta área são melhorar a segurança do paciente e melhorar os cuidados de saúde oferecidos, estes objetivos são compridos através da utilização de dados dos RES para investigação e nos procedimentos clínicos.

A chave do trabalho, e também uma parte importante para qualquer sistema de prospeção de dados e que use RES, é ter a habilidade de transformar dados heterogéneos dos pacientes, guardados em várias clínicas e em sistemas de saúde, em dados padronizados que depois podem ser comparados e pesquisados.

É uma plataforma que se foca em 4 áreas:

1. Modelação de informação clínica e padrões de terminologias;

2. Uma plataforma para a normalização de padronização de dados clínicos estruturados e não estruturados dos RES;

3. Uma plataforma para representar e executar a identificação de um grupo de pacientes e a fenotipagem lógica;

4. Avaliar a qualidade e utilidade dos dados usados pelo SHARPn;

A plataforma desenvolvida pela SHARPn foi construída usando CEMs (Clinical Elements Models), serviços de processamento de linguagem natural, serviços de motores de busca de terminologias (cTAKES) e serviços de terminologias comum (CTS2).

Para que os sistemas de suporte à decisão e os sistemas analíticos funcionem, os dados dentro dos RES devem ser normalizados. Sabendo isto, os autores desenvolveram vários métodos para a plataforma. A plataforma é capaz de normalizar e padronizar informação clínica e no caso da informação ser no formato de narrativo esta terá que ser processada por algoritmos de linguagem natural. É também capaz de transferir, guardar, processar e devolver dados (este processo consiste em fazer perguntas à base de dados). Outra das características da plataforma é identificar e extrair fenótipos (características visíveis de um individuo) presentes nos RES. A plataforma foi desenvolvida para assegurar a qualidade e a consistência dos dados produzidos.

Para testar o processo de identificação de fenótipos, os autores decidiram usar a medida NQF 0064, que determina a percentagem de pacientes entre os 18 e os 75 com diabetes e com um nível de LDL-C menor a 100 mg/dl. Onde o critério denominador identifica pacientes entre os 18 e os 75 anos que tenham sido diagnosticados com diabetes e o critério numerador identifica pacientes com níveis de LDL-C menor a 100 ml/dl. Foram usados dados de 273 pacientes, após a plataforma normalizar os dados foi usada a medida NQF 0064, que identificou 21 pacientes para o denominador (pacientes diagnosticados com diabetes) e 18 para o numerador (pacientes com níveis de LDL-C inferior a 100 ml/dl).

Capítulo 3

Desenho

Actualmente os sistemas de informação hospitalares utilizam dezenas de vocabulários diferentes para descrever os registos electrónicos médicos. Esta heterogeneidade torna difícil não só a integração de informação entre diferentes unidades de saúde, mas até mesmo a integração da informação dentro da mesma unidade. Muitos destes vocabulários são apenas uma lista de termos textuais controlada, sem qualquer tipo de informação sobre o significado ou relações com outros termos associada. Uma possível solução para este problema passa por utilizar ontologias de maior riqueza semântica para descrever os dados clínicos, que facilitem a sua integração e análise. Para tal é necessário desenvolver estratégias capazes de analisar os dados de maneira a descobrir qual ou quais as ontologias mais adequadas para os descrever e que faça também a anotação dos dados, escolhendo qual o melhor conceito para anotar determinado termo.

Tendo este objetivo em mente, foi desenvolvida uma estratégia baseada num sistema composto por três módulos: o **ExportConcepts**, o **RecommendOntologies** e o **EHRannotator**, que se articulam para cumprir o objectivo.

O módulo **exportConcepts** é utilizado para analisar os conceitos de uma ontologia e descobrir informação semântica sobre estes, para isso recebe uma ontologia e calcula uma série de valores para cada conceito na ontologia e que serão utilizados noutro módulo. As métricas utilizadas para avaliar os conceitos neste módulo são a centralidade, a especificidade e a densidade, que serão explicadas no próximo capítulo. No entanto, o sistema pode ser modificado para incluir outras métricas se necessário. Estas métricas são úteis para descobrir qual o melhor conceito para anotar um determinado termo e assim cumprir com um dos objetivos propostos ao mesmo tempo que se resolvem outros problemas associados com as anotações, como escolher o melhor conceito dentro de um conjunto de conceitos pertencentes a diferentes ontologias.

O módulo **RecommendOntologies** tem como propósito descobrir qual a melhor ontologia, ou melhores ontologias a utilizar para anotar um conjunto de termos clínicos. Neste trabalho, pretendia-se estudar a aplicação e adaptação de técnicas de recomendação em três diferentes contextos, global, paciente e evento. Estes três níveis apresentam desafios diferentes, pois correspondem progressivamente a um estreitamento dos domínios. Ao fazer recomendações a nível global, temos apenas uma ontologia ou conjunto de ontologias a aplicar a todo o conjunto de dados. No entanto, ao fazer a recomendação a nível do paciente ou evento, teremos um número cada vez mais elevado de recomendações que será depois necessário analisar e convergir no conjunto óptimo. É também relevante que o sistema consiga recomendar mais que uma ontologia para garantir uma boa cobertura dos vários domínios envolvidos.

Existem várias estratégias na literatura que se focam sobre a recomendação de ontologias como a ferramenta Recommender (ferramenta do BioPortal) [14] [20], o sistema AKTiveRank [12], que avalia as ontologias baseando-se nas suas estruturas e na forma como cobrem os termos de *input*, outra estratégia foi desenvolvida para avaliar ontologias tendo em conta a forma como a ontologia cobre um determinado contexto, a sua riqueza semântica para esse contexto e a popularidade da ontologia [21].

O sistema pode integrar neste módulo diferentes estratégias, desde que estas recebam como *input* um conjunto de termos em texto, e devolvam o conjunto de ontologias (de 1 a 3) mais indicadas.

Para este trabalho foi utilizado o *webservice* de recomendação de ontologias do BioPortal, o Recommender, porque cumpre os requisitos necessários e porque está ligado ao maior repositório de ontologias biomédicas, sem ser preciso o esforço de reunir um catálogo de ontologias. O Recommender tem um funcionamento bastante simples, recebe como *input* texto ou um conjunto de palavras-chaves e como *output*, dependendo da configuração, devolve um conjunto de ontologias, que foram escolhidas pelo sistema como sendo as mais adequadas para o *input*.

O módulo **EHRannotator** tem como objectivo anotar os termos clínicos com as ontologias e seus conceitos mais adequados. Este módulo é responsável pela gestão e integração dos resultados dos outros dois módulos, o RecommendOntologies e o ExportConcepts. O EHRannotator recebe como *input* uma série de parâmetros definidos pelo utilizador: o número de ontologias a utilizar (entre 1 a 3 ontologias), o tipo de *input* para os termos (“texto” ou “palavras-chave”) e a forma de filtragem (valor de cobertura da ontologia, centralidade do conceito, especificidade do conceito e densidade do conceito). Os valores de filtragem serão importantes para casos em que existem vários conceitos associados a um termo. Neste caso, é preciso escolher o melhor conceito, o

conceito com maior valor para o critério escolhido pelo utilizador é o escolhido. O critério de filtragem será escolhido previamente pelo utilizador. O *output* do módulo corresponde à melhor anotação identificada para cada termo, utilizando apenas as ontologias recomendadas.

A combinação destes 3 módulos, exemplificada na Figura 3.1, cumpre com sucesso os 2 objetivos propostos no início do projeto, que eram os seguintes:

- Descobrir a melhor ontologia a usar para o determinado termo;
- Descobrir a melhor anotação de um conjunto de anotações para um determinando termo;

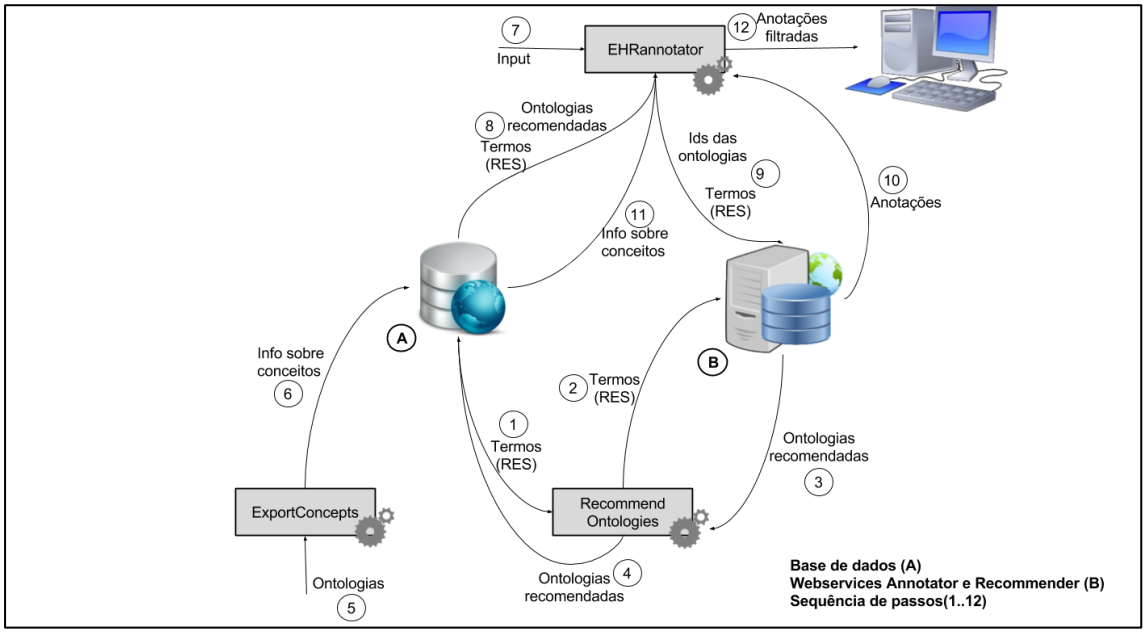

Figura 3.1- Desenho do sistema com a representação dos módulos RecommendOntologies, ExportConcepts , EHRannotator, base de dados, webservices e fluxo de dados;

Capítulo 4

Implementação

Neste capítulo é descrita em detalhe a implementação de cada módulo do sistema, bem como as novas contribuições atingidas com este trabalho.

Antes de explicar como é que foi implementado cada módulo é importa saber quais as tecnologias utilizadas. Para os módulos foi utilizada a linguagem JAVA e para a base de dados foi utilizada a linguagem MySQL. Para a realização das tarefas dos módulos foram utilizadas as seguintes bibliotecas JAVA: JDBC, para comunicar com a base de dados; JENA, para navegar dentro das ontologias e a biblioteca Jackson, usada para extrair dados das respostas retornadas pelos *webservices*.

4.1 ExportConcepts

O ExportConcepts, primeiro carrega a ontologia e depois com ajuda da API Jena [23], analisa cada conceito da ontologia. Esta análise tem em conta a localização do conceito dentro da ontologia e o número de descendentes do conceito para calcular os valores de centralidade, especificidade e densidade. Estes valores serão depois combinados com as anotações criadas pelo EHRannotator.

A forma utilizada para normalizar as pontuações de cada conceito é baseada na estrutura da ontologia donde este conceito foi retirado, ou seja, não tem em conta outras ontologias. Contudo é a forma mais justa para comparar conceitos de diferentes ontologias. Acrescenta alguma informação sobre a qualidade do conceito à anotação.

A cada conceito são atribuídos 3 valores diferentes, valor de centralidade, valor de especificidade e o valor de densidade. O valor de centralidade e densidade foram inspirados no artigo [16]. Juntamente com o valor de especificidade, estes valores serão úteis para o processo de filtragem, que será explicado no capítulo EHRannotator.

Estes valores foram escolhidos porque para além de facilitarem o processo de escolher um conceito para um termo, apresentam, de uma forma simples, a qualidade semântica que um conceito tem dentro da ontologia a que pertence.

Para calcular o valor de centralidade de um conceito, é necessário descobrir todos os caminhos que comecem na raiz e terminem numa folha e que passem por esse conceito, e quais desses caminhos é que têm o conceito, que se está a estudar, no meio desses caminhos. Este critério é útil para escolher termos que não sejam muito específicos, nem muito superficiais. Critério que identifica se um determinado conceito podia servir como categoria natural. O exemplo que se segue foi retirado do artigo [16] e explica o que é um termo central e como pode ser útil. De um conjunto de termos, por exemplo, "veículo", "carro" e "carro desportivo", o termo que identifica melhor uma família de veículos é o termo "carro" porque não é tão específico como o termo "carro desportivo", nem tão geral como "veículo". No caso das anotações, esta pontuação é útil caso o utilizador pretenda anotar os seus dados com termos mais centrais, termos que não são muito específicos, nem muito gerais.

A fórmula utilizada para calcular o valor centralidade é a seguinte:

$$C(C,O) = \frac{Cmc(C)}{TCm(C)}$$

Equação 4.1- Fórmula da centralidade

Onde *Cmc* são todos os caminhos que têm o conceito *C* no meio e o *TCm* são todos os caminhos que passam pelo conceito *C*.

O valor de especificidade é calculado através da profundidade do conceito e da profundidade máxima da ontologia. Através deste valor é possível perceber como é que a ontologia, de onde o conceito foi retirado, especifica o conceito através das suas relações com os seus antepassados.

$$E(C,O) = \frac{P(C)}{Pm(O)}$$

Equação 4.2- Fórmula da especificidade

Onde *P* é a profundidade onde o conceito se encontra e *Pm* é a profundidade máxima da ontologia. Este valor é útil se o utilizador quiser anotar os seus termos com conceitos mais específicos.

A última pontuação é o valor de densidade, é calculada usando o número de descendentes do conceito e o número de descendentes máximo da ontologia. Este valor representa a forma como aquela ontologia descreve aquele conceito usando os descendentes deste. Se um conceito tiver poucos ou nenhuns descendentes, é um conceito com pouca informação. O mesmo não acontece com um conceito com muitos descendentes, sendo este bem descrito pela ontologia.

$$D(C,O) = \frac{Dsc(C)}{DscM(O)}$$

Equação 4.3- Fórmula da densidade

Onde *Dsc* é o número de descendentes no conceito *C* e o *DscM* o número de descendentes máximos dentro da ontologia.

Para perceber melhor estes valores será explicado como é que estes valores são calculados usando o grafo exemplo que está presente na Figura 4.1. O conceito a ser avaliado será o C1.

Há valores que se tem de ter em conta para calcular a especificidade e a densidade do conceito, não só os valores diretamente relacionados com o conceito, mas outros valores, como a *profundidade máxima da ontologia* e *número máximo de descendentes da ontologia*. Para o caso do grafo da Figura 4.1, a profundidade máxima é 5 e o número máximo de descendentes num conceito é 4.

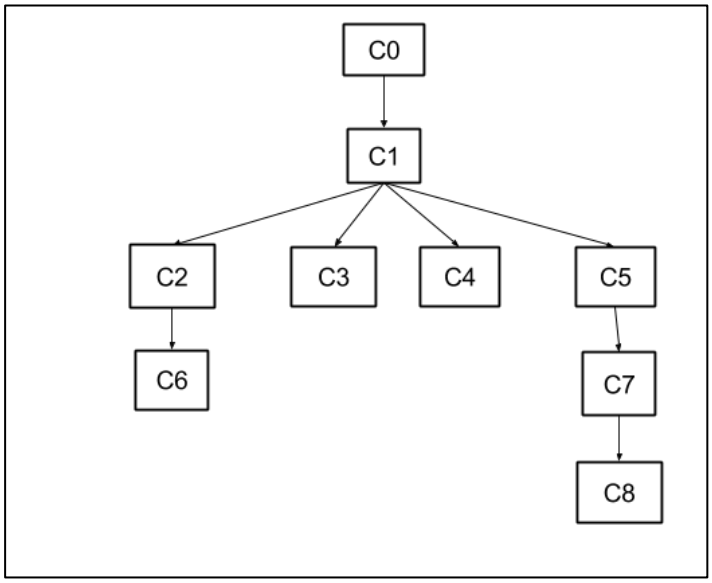

Figura 4.1- Grafo exemplo

Como é possível ver através do esquema, o conceito C1 tem 4 caminhos a passar por ele, em que 3 destes caminhos o conceito C1 está no meio, o que faz com que a sua centralidade seja de 0.75. É importante saber que um conceito não tem que estar exactamente no meio de um caminho para que esse caminho seja considerado como um caminho onde conceito está no meio. Por exemplo, quando o caminho tem um comprimento par, o conceito tem de estar num dos nós do meio para que esse caminho seja selecionado e considerado como um caminho onde o conceito está no meio, que é o que acontece no caminho C0-C1-C2-C6. É claro que quando o caminho tem comprimento ímpar o conceito é obrigado a estar no meio para que o caminho seja considerado como um caminho onde o conceito está no meio. A especificidade do conceito é de 0.4, pois a profundidade do conceito é 2 e a total é de 5. O valor de densidade é calculado com o número descendentes, que para este conceito são 4, e é usado o número de conceitos máximo da ontologia, que também são 4, sendo assim o valor de densidade para este conceito é de 1.

4.2 RecommendOntologies

O RecommendOntologies foi utilizado para fazer uma análise a cada agrupamento de conceitos em todas as situações possíveis, ou seja, analisar os termos como "texto" ou "palavras-chaves" e recomendar ontologias, e guardar as recomendações na base de dados. É com estes dados que depois se descobre qual a melhor ontologia ou ontologias a usar para fazer a anotação de um determinado agrupamento de termos.

4.2.1 Agrupamento de termos

Tipicamente, as ferramentas de anotação semântica têm um tamanho máximo (palavras ou termos) que conseguem processar. Os vocabulários clínicos são tipicamente de tamanho considerável (milhares de termos), a estratégia proposta considera três estratégias para agrupamentos de termos (paciente, evento, global), que permitem a utilização do serviço do BioPortal para anotar a totalidade dos termos clínicos. Outros aspetos podem ser analisados como a cobertura das ontologias recomendadas. No início, o objectivo do agrupamento "global" era juntar todos o termos utilizados nos registos electrónicos de saúde do Openmrs e fazer um pedido de recomendação ao Recommender para mais tarde comparar os resultados com o dos outros agrupamentos (paciente e evento). Basicamente, a ideia era confirmar se era

necessário fazer uma análise prévia a cada paciente e a cada evento, ou simplesmente analisar de uma só vez todos os termos utilizados na base de dados. Contudo esta ideia não pôde ser concretizada devido à incapacidade dos *webservices* (Annotator, Recommender) em processar muitos termos de uma vez. Para contornar este problema criou-se agrupamentos de 75 termos, de forma a recolher recomendações para o contexto global.

O outro agrupamento é o paciente. Neste caso os termos são agrupados por um determinado paciente e enviados ao Recommender para que este faça a recomendação. O último agrupamento tem os conceitos agrupados por evento, todos os conceitos relacionados com um determinado evento são enviados ao Recommender para que este faça a recomendação.

4.2.2 Funcionamento do RecommendOntologies e selecção de ontologias

O RecommendOntologies faz pedidos ao *webservice* Recommender, sendo estes pedidos constituídos por termos (*input*), técnica de anotação para a análise ("texto" ou "palavra-chave") e número ontologias a recomendar (lista com ontologias ou conjuntos de ontologias). O Recommender envia uma resposta em JSON, onde vêm as ontologias recomendadas. Desta resposta é extraída a melhor ontologia ou o melhor conjunto de ontologias recomendadas, juntamente com as respectivas avaliações. Estas informações, como a técnica usada na análise, agrupamento (global, paciente, ou evento), número de ontologias recomendadas e ontologias, são guardadas na base de dados para que depois possam ser usadas pelo EHRannotator.

É importante explicar o processo de seleção de ontologias. O *webservice* Recommender, quando envia uma resposta, independentemente se é uma lista de conjuntos de ontologias ou uma lista de ontologias, estas vêm ordenadas através de uma pontuação final que é calculada através da avaliação da cobertura, aceitação, detalhe e especificação. É importante salientar que estas pontuações não são todas fixas e que podem mudar dependendo dos termos enviados. De todas as ontologias que são enviadas pelo Recommender, e para cada caso, apenas é guardada a ontologia que vem em primeiro lugar. A Figura 4.2 demonstra os principais processos do RecommendOntologies.

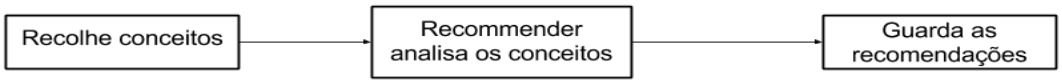

Figura 4.2- Principais processos do RecommenderOntologies

O módulo, recolhe os conceitos por “Global”, “Paciente” ou “encontro”, envia-os para o Recommender juntamente com outros parâmetros e recebe as recomendações. Dessas recomendações apenas é guardada a que está em primeiro lugar. Quanto maior for o número de agrupamentos utilizados, maior será o número de recomendações obtidas, o que irá facilitar a tarefa de escolher qual a ontologia ou ontologias a utilizar para fazer a anotação.

4.3 EHRannotator

Este módulo vai combinar os dados obtidos pelos módulos anteriores, ou seja, os valores dos conceitos e as recomendações, para obter uma lista de anotações filtrada da forma que o utilizador definir.

O esquema de funcionamento é semelhante ao do RecommenderOntologies, mas com algumas diferenças. O módulo não agrupa os conceitos por “global”, este agrupamento de termos não foi utilizado porque a ideia era usar as recomendações deste agrupamento para compará-las com as recomendações dos outros agrupamentos, por isso é que não são feitas anotações para o agrupamento “global”. As outras diferenças devem-se ao facto deste programa fazer anotações e não recomendações, logo terá funcionalidades úteis para cumprir este objectivo, como usar o Annotator e filtrar anotações.

Mais uma vez utilizou-se uma ferramenta do BioPortal, em vez de se criar uma, foi utilizado o *webservice* Annotator, visto que o foco deste trabalho não era desenvolver uma ferramenta de anotação mas melhorar a forma como as anotações são feitas.

No EHRannotator, o utilizador insere o número de ontologias, o tipo de filtro, e a forma como os termos devem ser analisados. Sabendo isto o programa escolhe a ontologia ou ontologias a serem usadas para fazer a anotação dos termos. A técnica de análise, termos e ontologias recomendadas são enviados (pelo EHRannotator) para o Annotator, este responde com uma lista de anotações que depois irá ser filtrada. A Figura 4.3 representa os principais processos do EHRannotator.

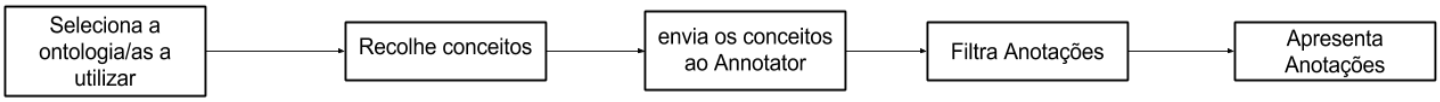

Figura 4.3- Processos principais do EHRannotator

4.3.1 Processo de filtragem

Durante o desenvolvimento do relatório preliminar, e também através de testes feitos com o *webservice* Annotator, verificou-se que existiam alguns possíveis problemas a serem resolvidos no processo da anotação, tais como:

- Um termo ser anotado com várias classes da mesma ontologia;
- Um termo ser anotado com classes de diferentes ontologias;

Antes de se iniciar o processo de filtragem há um passo importante para que isto se possa realizar, que é aplicar uma pontuação a cada anotação. Esta pontuação depende do tipo de filtro (critério) que o utilizador escolheu. Podem ser escolhidos 3 tipos de filtros diferentes:

- Centralidade – a avaliação da anotação vai ser igual ao valor de centralidade do conceito ao qual está ligada;
- Especificidade – a avaliação da anotação vai ser igual ao valor de especificidade do conceito ao qual está ligada;
- Densidade – a avaliação da anotação vai ser igual ao valor de densidade do conceito ao qual está ligada;

Para perceber melhor o processo de filtragem e como é que este soluciona os problemas mencionados anteriormente, analisemos o termo “diagnosis”, que está contido em “diagnosis added” e está associado ao paciente com o id 2. É importante saber que os termos deste caso foram analisados com a técnica “texto”, o número de ontologias utilizadas foram 3 e que o filtro usado foi “especificidade”.

A Tabela 4.1 mostra todos os conceitos utilizados para anotar o termo "diagnosis" antes da filtragem, ou seja, antes da escolha do melhor conceito para aquele termo:

Classe (conceito)	Ontologia	Valor de cent.	Valor de esp.	Valor de den.
C0011900	HL7	0	0,6	0
C49653	NCIT	0	0,4375	0,00030911
29308-4	LOINC	0	0,25	0
MTHU008876	LOINC	0	0,0625	0
C15220	NCIT	0	0,3125	0,001545
LP72437-4	LOINC	0	0,185	0

Tabela 4.1- Tabela com conceitos e valores das anotações

Sabendo que o utilizador escolheu o filtro "especificidade", o conceito que irá ser usado para anotar o termo "diagnosis" é o conceito C0011900 da ontologia HL7 porque de todos os conceitos é o conceito com maior valor de especificidade.

4.3.2 Cálculo da cobertura para cada processo de anotação

No final de cada processo de anotação, o EHRannotator calcula a percentagem de cobertura da ontologia ou ontologias sobre o *input*. O processo de calcular a cobertura difere um pouco dependendo da forma como os termos são analisados ("texto" ou "palavra-chave"). As anotações usadas para calcular o valor da cobertura são as que permanecem após o processo de filtragem.

O processo consiste em comparar cada termo do *input* com cada conceito de todas as anotações. No caso em que a análise aos termos de *input* é feita por "texto", existe maior liberdade em comparar os termos com os conceitos pois um termo formado por várias palavras pode ser comparado de diferentes formas com os conceitos, o mesmo não acontece quando os termos são analisados como "palavras-chave". Neste caso, em que a análise é feita através por "palavras-chave", os termos são comparados com os conceitos de todas as anotações, mas a comparação é feita com a totalidade do termo, não existe a liberdade de dividir o termo em vários termos e comparar o conceito com cada um dos termos.

Para explicar melhor a forma como é calculado o valor de cobertura vejamos o *input* "blood cell type" e "stomach". Para este caso, quando se usa a ontologia NCIT as anotações resultantes são as que estão na Tabela 4.2.

CLASS filter	ONTOLOGY filter	TYPE filter	CONTEXT	MATCHED CLASS filter	MATCHED ONTOLOGY filter
Peripheral Blood Cell	National Cancer Institute Thesaurus	direct	**blood cell** type, stomach	Peripheral Blood Cell	National Cancer Institute Thesaurus
Mouse Blood	National Cancer Institute Thesaurus	direct	**blood** cell type, stomach	Mouse Blood	National Cancer Institute Thesaurus
Blood	National Cancer Institute Thesaurus	direct	**blood** cell type, stomach	Blood	National Cancer Institute Thesaurus
Cell	National Cancer Institute Thesaurus	direct	blood **cell type**, stomach	Cell	National Cancer Institute Thesaurus
Cell	National Cancer Institute Thesaurus	direct	blood **cell** type, stomach	Cell	National Cancer Institute Thesaurus
Cell Device Component	National Cancer Institute Thesaurus	direct	blood **cell** type, stomach	Cell Device Component	National Cancer Institute Thesaurus
Cell	National Cancer Institute Thesaurus	direct	blood **cell** type, stomach	Cell	National Cancer Institute Thesaurus
Cellular Telephone	National Cancer Institute Thesaurus	direct	blood **cell** type, stomach	Cellular Telephone	National Cancer Institute Thesaurus
Type	National Cancer Institute Thesaurus	direct	blood cell **type**, stomach	Type	National Cancer Institute Thesaurus
Stomach	National Cancer Institute Thesaurus	direct	... cell type, **stomach**	Stomach	National Cancer Institute Thesaurus
Gastric Tissue	National Cancer Institute Thesaurus	direct	... cell type, **stomach**	Gastric Tissue	National Cancer Institute Thesaurus
Mouse Stomach	National Cancer Institute Thesaurus	direct	... cell type, **stomach**	Mouse Stomach	National Cancer Institute Thesaurus

Tabela 4.2- Anotações para o termo "blood cell type, stomach"

A pontuação de cobertura para a anotação tipo "texto" seria de 100%, pois cada palavra do termo "blood cell type" e "stomach" pode ser anotada usando os conceitos que estão na Tabela 4.2. Para o tipo "palavras-chaves", em que o conceito tem que ser igual à totalidade do termo, a cobertura é de 50%, pois o único termo coberto na totalidade por um conceito é o termo "stomach".

A fórmula para calcular a percentagem de cobertura para o termos inseridos é:

$$Percentagem\ de\ cobertura = \frac{\sum valor\ de\ cobertura\ do\ termo}{número\ de\ termos}$$

Equação 4.4 - Valor de cobertura calculado pelo EHRannotator

4.3.3 Diferentes fórmulas para calcular o valor de cobertura

A pontuação de cobertura do serviço do Bioportal, calculada pelo Recommender, é obtida através de uma fórmula diferente da que foi utilizada para calcular a cobertura das ontologias sobre os termos de *input* no programa EHRannotator.

A pontuação da cobertura calculada pelo Recommender tem em conta as anotações obtidas pelo Annotator e nunca o número total de termos de *input* ou o número de termos não anotados. A cada uma das anotações é da uma pontuação, essa pontuação é calculada multiplicando os factores pelo número de palavras da anotação.

Os factores são os seguintes:

- 10 – Se a anotação for do tipo "pref", ou seja, o nome do conceito é igual ao termo;
- 6 – Se a anotação for formada por várias palavras;
- 5 – Se a anotação for associada a um sinónimo do conceito, ou seja, o termo de *input* combina com o sinónimo do conceito;

No final, para normalizar a pontuação de cada ontologia, cada pontuação é dividida pela pontuação total. Esta pontuação total é calculada através da soma das pontuações das melhores anotações resultantes de se usarem todas as ontologias para anotar o *input*. As melhores anotações são escolhidas através de um processo de filtragem que rejeita anotações de sinónimos, anotações com pontuações baixas e anotações repetidas. Logo obtém-se a razão entre a pontuação obtida usando as ontologias escolhidas face à pontuação obtida usando todas as ontologias. A fórmula utilizada é a seguinte:

$$valor\ cb(O) = \frac{soma\ das\ pontuações(O)}{pontuação\ total(todas\ as\ ontologias)}$$

Equação 4.5 - Equação usada pelo Recommender para o valor de cobertura

Para perceber a diferença serão apresentados 3 valores de cobertura, um calculado pelo Recommender, outro é a média de todos os valores obtidos pelo Recommender guardados na base de dados e por último, o valor de cobertura calculado pelo EHRannotator.

Como a diferença entre os valores de cobertura é maior quando os termos são analisados como "palavras-chave", o caso escolhido será o encontro 26999 do paciente com o id igual a 2 e o termos foram analisados como "palavras-chave".

A Tabela 4.3 apresenta as ontologias recomendadas para esse caso (conjunto de uma ontologia e análise "palavras-chave"), é possível ver o valor de cobertura (coverage score) para cada ontologia.

POS.	ONTOLOGY	FINAL SCORE	COVERAGE SCORE	ACCEPTANCE SCORE	DETAIL SCORE	SPECIALIZATION SCORE	ANNOTATIONS
1	NCIT	73.6	81.0	88.5	77.6	27.7	13
2	SNOMEDCT	56.6	55.9	96.8	61.1	14.3	6
3	RCD	53.5	57.5	87.3	45.8	12.6	5
4	LOINC	48.1	50.8	89.9	28.6	15.8	7
5	EFO	42.7	42.2	35.3	83.8	11.1	4
6	NLMVS	42.2	40.6	23.4	9.2	100.0	3
7	MEDDRA	41.1	40.6	97.0	20.3	7.6	3
8	TRAK	39.5	45.4	19.7	59.9	17.3	6
9	CCONT	39.5	40.6	26.2	78.4	10.1	3
10	MESH	39.1	20.3	88.9	92.6	4.9	3

Tabela 4.3-Recomendações para o evento 26999 (uma ontologia, análise "palavra-chave")

Como é possível ver na Tabela 4.3, o Recommender deu à ontologia NCIT um valor de cobertura de 81%, foi assim calculado porque as suas anotações tiveram uma pontuação de 255, que a dividir por 318, que é a pontuação total das anotações usando todas as ontologias, obtém-se o valor normalizado de 0.81. O valor de cobertura média dos valores guardados na base de dados, foi de 80.2%, o que não é muito diferente do valor de cobertura calculado pelo Recommender para o caso em específico.

No caso do EHRannotator, o valor de cobertura é muito diferente dos anteriores, visto que é calculado de uma forma diferente. Para este caso, o EHRannotator, com as anotações que obteve do Annotator, calculou um valor de cobertura de 23.6%. Este valor não reflecte a qualidade das anotações, como no caso do valor calculado pelo Recommender, mas sim a forma como a ontologia usada cobre os termos de *input* com conceitos. Percebendo esta diferença de obter valores de cobertura através de fórmulas diferentes e que estas reflectem diferentes formas como uma ontologia ou conjunto de ontologias se comportam com um determinado *input*, torna-se mais fácil interpretar os resultados obtidos.

Capítulo 5

Resultados e Discussão

Neste capítulo serão analisados os dados obtidos pelo RecommendOntologies, e algumas anotações feitas a pacientes e encontros extraídos do EHRannotator (este dados encontram-se no anexo A e B), utilizando dados de uma base de dados de registos electrónicos médicos *open source*. Irei comparar a semelhança entre as ontologias recomendadas para cada um dos agrupamentos, em cada uma das diferentes situações. Será discutida a utilidade e aplicação das diferentes formas de agrupar os termos (“paciente” ou “evento”), e as vantagens e desvantagens de ambas as formas de analisar os dados (“texto” ou “palavras-chave”), bem como qual a melhor forma para descobrir que ontologia ou ontologias a usar para anotar os dados.

No término do programa RecommendOntologies a base de dados tinha 111843 recomendações, 287 para agrupamento global, 31668 para o agrupamento dos pacientes e 79888 para o agrupamento dos eventos. Com esta quantidade de recomendações torna-se possível encontrar a ontologia ou ontologias a usar para anotar os termos.

5.1 Fonte de dados

A amostra de registos eletrónicos de saúde utilizada para popular a base de dados foi criada pela Openmrs [23], uma plataforma clínica aberta de registos médicos criada para ajudar países em desenvolvimento. A base de dados contém 5000 pacientes, 15000 encontros e 500000 observações.

A Figura 5.1 representa um excerto da base de dados da Openmrs, contemplando apenas as tabelas que foram usadas.

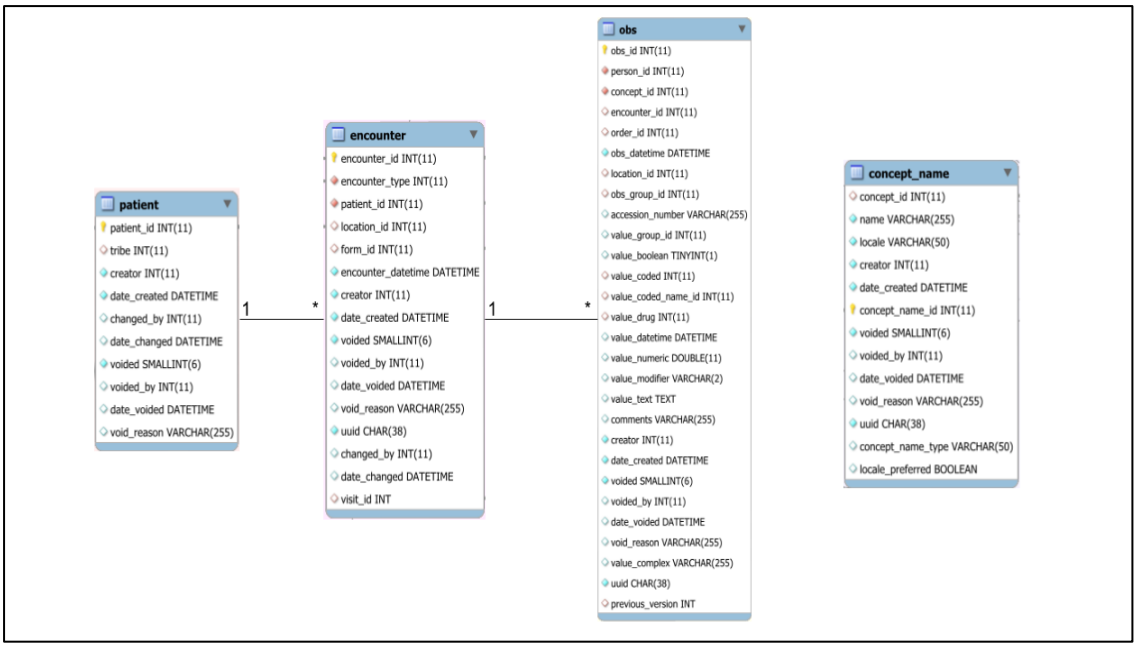

Figura 5.1 - Tabelas do Openmrs

A ideia é agrupar todos os conceitos por encounter (evento) e patient (paciente). Como é possível observar, cada encounter tem várias obs, cada uma delas com um concept_id. São estes conceitos que serão agrupados por "global", "paciente" e "evento", para mais tarde serem utilizados como *input* para se fazerem as recomendações e as anotações.

5.2 Configurações

Como é possível pedir ao Recommender para recomendar uma ou mais ontologias e analisar os conceitos como "palavras-chaves" ou "texto", o RecommendOntologies pode ser configurado de diferentes formas, o que resultou em vários casos de estudo diferentes. A combinação dos parâmetros de tipo de análise, agrupamento e número de ontologias, originou os seguintes casos:

- Global, texto, 1 ontologia;
- Global, texto, 2 ontologias;
- Global, texto, 3 ontologias;
- Global, palavra-chave, 1 ontologia;
- Global, palavra-chave, 2 ontologias;
- Global, palavra-chave, 3 ontologias;

- Paciente, texto, 1 ontologia;
- Paciente, texto, 2 ontologias;
- Paciente, texto, 3 ontologias;
- Paciente, palavra-chave, 1 ontologia;
- Paciente, palavra-chave, 2 ontologias;
- Paciente, palavra-chave, 3 ontologias;
- Evento, texto, 1 ontologia;
- Evento, texto, 2 ontologias;
- Evento, texto, 3 ontologias;
- Evento, palavra-chave, 1 ontologia;
- Evento, palavra-chave, 2 ontologias;
- Evento, palavra-chave, 3 ontologias;

5.3 Resultados do RecommendOntologies

Como foi apresentado no capítulo de implementação, foram criados 3 agrupamentos diferentes, e cada agrupamento de termos foi analisado pelo RecommenderOntologies de duas formas diferentes, como “palavras-chaves” ou “texto”, para recolher a melhor ontologia recomendada ou o melhor conjunto de ontologias.

Os resultados do RecommendOntologies são constituídos por dois valores para cada ontologia ou conjuntos de ontologias: valor de cobertura médio (cov) e avaliação ponderada (weighted_ava). O valor de cobertura médio reflecte a qualidade das anotações obtidas pelas ontologias recomendadas e a avaliação ponderada reflecte a qualidade da ontologia ou conjunto de ontologia em relação a outras ontologias ou conjuntos de ontologias guardadas na base de dados e tem em conta o número de vezes que foram recomendadas.

Para a situação em que se pediu ao RecommendOntologies para analisar os termos como "texto" e recolher a primeira ontologia recomendada pelo Recommender. As Figuras 5.2, 5.3 e 5.4 representam os resultados obtidos para cada um dos agrupamentos.

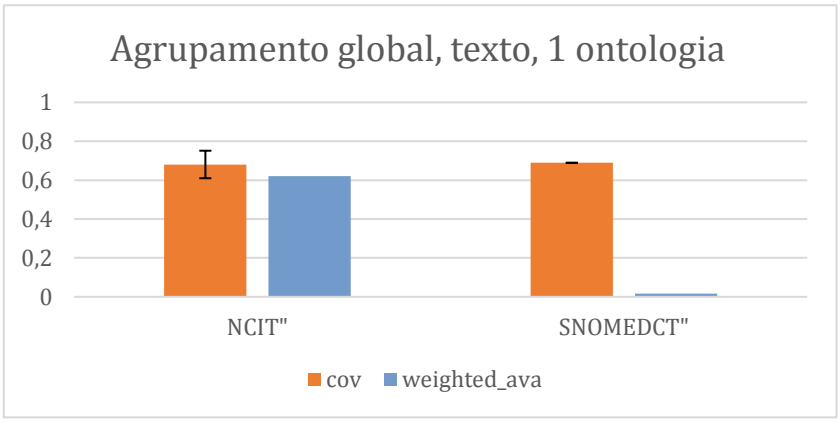

Figura 5.2 - Gráfico de recomendações para o agrupamento global, texto, 1 ontologia

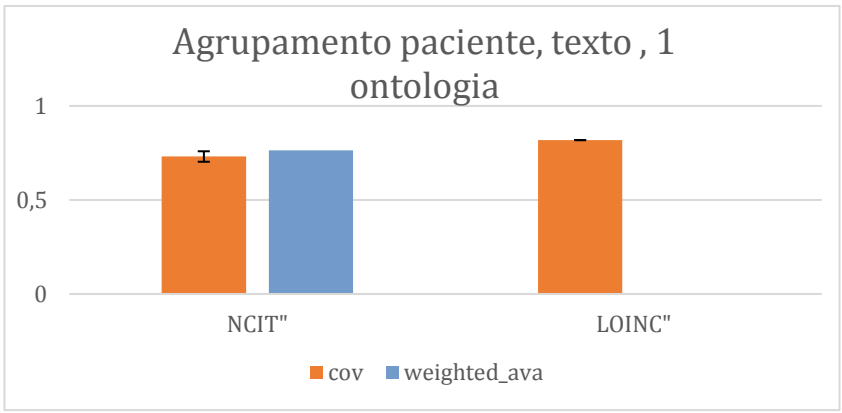

Figura 5.3 - Gráfico de recomendações para o agrupamento paciente, texto, 1 ontologia

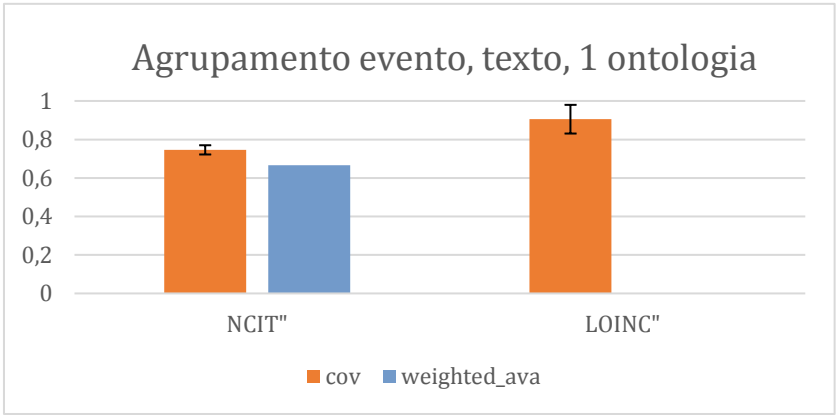

Figura 5.4- Gráfico de recomendações para o agrupamento evento, texto, 1 ontologia

Segundo os dados obtidos pelo programa, a ontologia recomendada, que aparece em primeiro, lugar é a NCIT para cada um dos agrupamentos, o que varia é a qualidade das anotações em cada um dos agrupamentos.

Os valores médios de cobertura para os agrupamentos "global", "paciente" e "evento" foram 0.68, 0.73, 0.74, respetivamente. O agrupamento com as melhores anotações foi o "evento", o que significa que se os termos forem agrupados por "evento" as anotações resultantes terão uma boa pontuação.

O valor de cobertura médio aumenta do agrupamento "global" para o agrupamento "paciente" e deste para o agrupamento "evento" porque os termos deixam de serem agrupados de uma forma aleatória, como acontece no "global", ou seja, estão mais relacionados entre si e pertencem quase todos ao mesmo domínio, o que consequentemente faz aumentar a pontuação das anotações e aumentar o valor de cobertura médio. Apesar de a ontologia em primeiro lugar ser uma ontologia mais genérica em todos os agrupamentos, as ontologias que vêm em segundo lugar para os agrupamentos "paciente" e "evento" são mais específicas, o que não acontece no agrupamento "global". Para o agrupamento "global", é a SNOMED-CT que é uma ontologia genérica e para os outros agrupamentos é a ontologia LOINC, que é uma ontologia mais específica.

O caso a ser analisado a seguir é aquele em que os termos foram analisados como "texto" e foi pedido ao Recommender para recomendar conjuntos de duas ontologias, as Figuras 5.5, 5.6 e 5.7 representam os resultados obtidos.

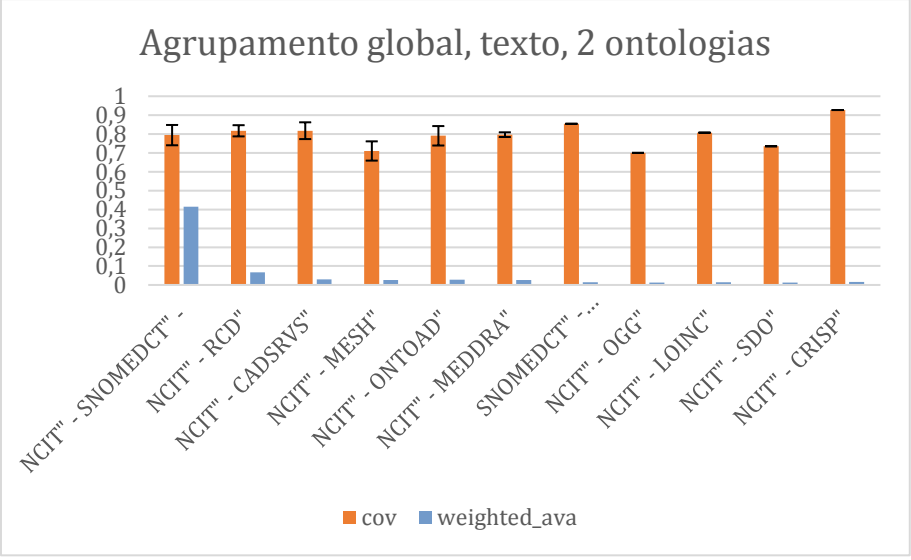

Figura 5.5- Gráfico de recomendações para o agrupamento global, texto, 2 ontologias

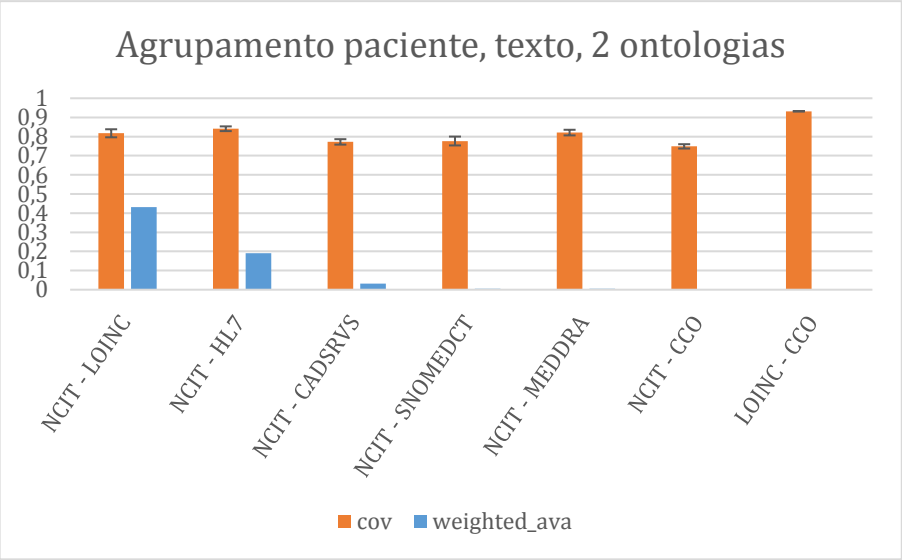

Figura 5.6 - Gráfico de recomendações para o agrupamento paciente, texto, 2 ontologias

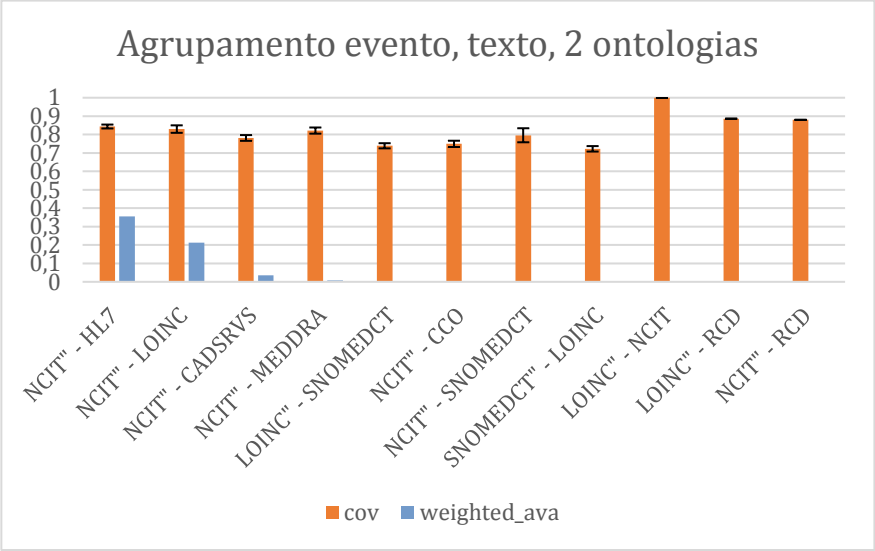

Figura 5.7 - Gráfico de recomendações para o agrupamento eventos, texto, 2 ontologias

Neste caso as recomendações foram um pouco diferentes. Contudo todos os conjuntos que estavam em primeiro lugar tinham a ontologia NCIT e mais uma vez o agrupamento com a melhor cobertura foi o "evento" com 0.84. Os outros agrupamentos "global" e "paciente" tiveram 0.79 e 0.81, respectivamente. Como é possível ver o valor de cobertura aumenta quanto maior for o número ontologias utilizadas.

Cada ontologia é diferente uma da outra, cobrem domínios diferentes e por isso quantas mais ontologias se usarem maior será o valor de cobertura. Isto também explica o porquê de cada agrupamento ter ontologias diferentes. Os termos dentro de cada agrupamento, apesar de pertencem ao mesmo vocabulário controlado, pertencem a domínios diferentes e este facto faz variar as ontologias recomendadas.

O próximo caso a ser analisado e representado pelas Figuras 5.8, 5.9 e 5.10, será o caso em que o termos foram analisados como "texto" e foi pedido ao Recommender para recomendar conjuntos de 3 ontologias.

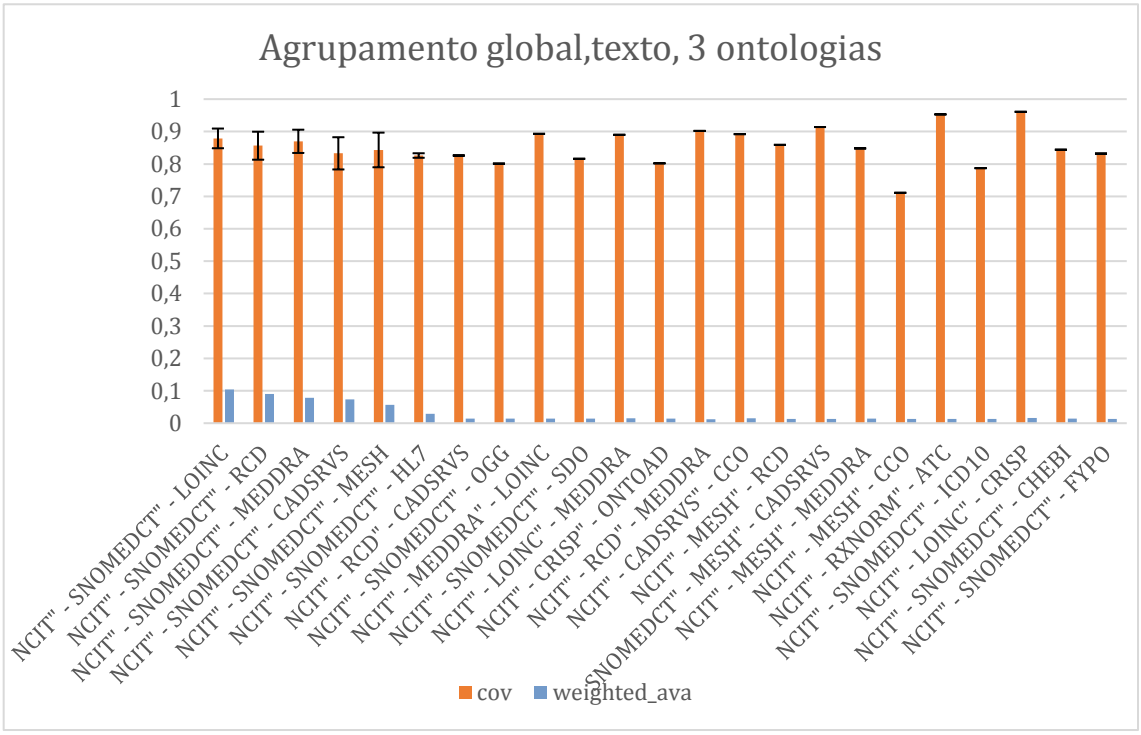

Figura 5.8- Gráfico de recomendações para o agrupamento global, texto, 3 ontologias

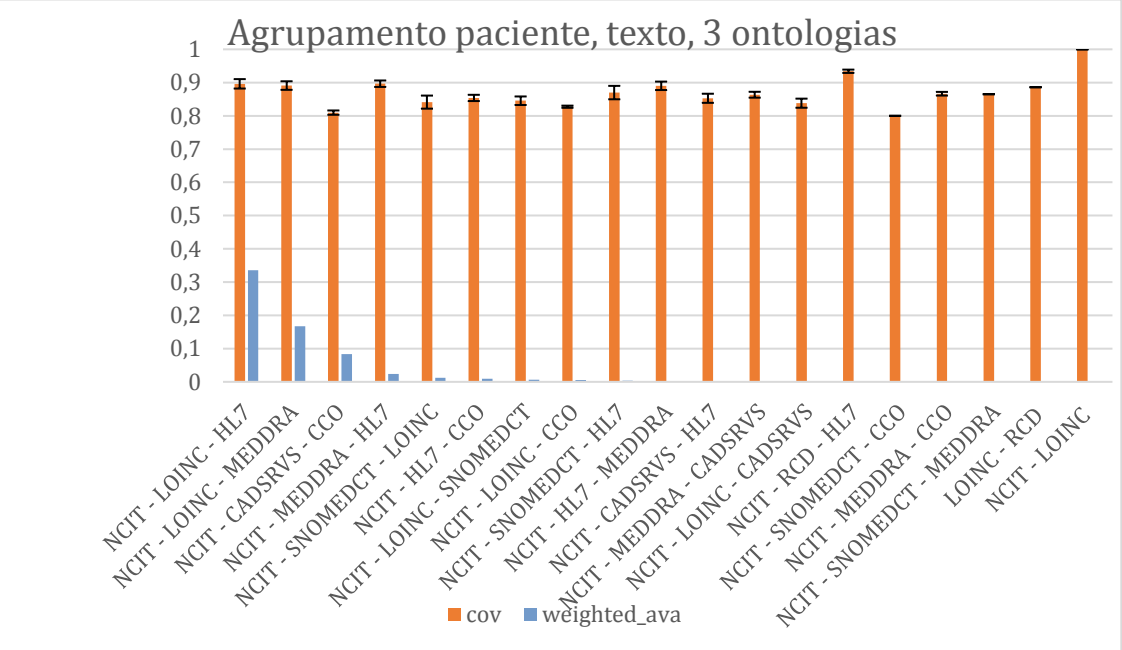

Figura 5.9- Gráfico de recomendações para o agrupamento paciente, texto, 3 ontologias

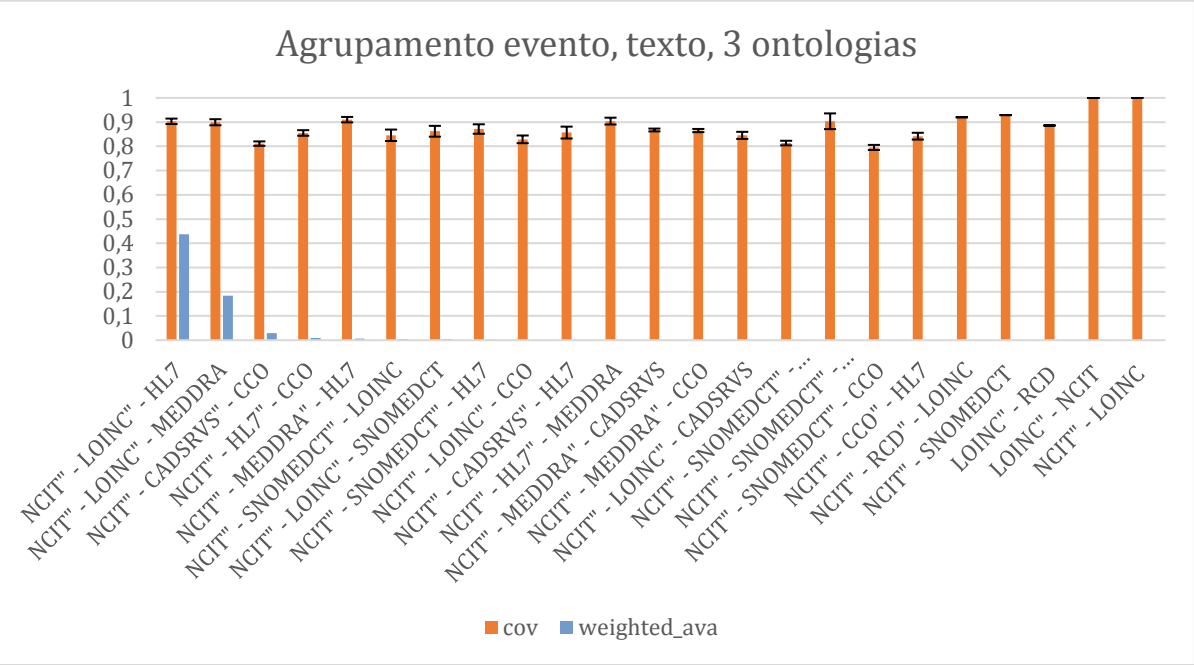

Figura 5.10 - Gráfico de recomendações para o agrupamento evento, texto, 3 ontologias

Como no caso anterior, houve um aumento do valor médio de cobertura em todos os agrupamentos, e mais uma vez, o agrupamento com o maior valor médio foi o "evento".

As ontologias recomendadas foram ligeiramente diferentes; o agrupamento "evento" e "paciente" tiverem em primeiro lugar o mesmo conjunto de ontologias e o agrupamento "global"; em relação aos outros dois, a única ontologia diferente é a SNOMED-CT.

Se os termos forem anotados como "texto" a melhor forma de os agrupar para garantir que estes são anotados com as melhores anotações é agrupá-los por "evento" e usar mais de uma ontologia para anotá-los. Esta é a forma que garante melhores resultados para o tipo "texto".

Os próximos casos a serem analisados e representados pelas Figuras 5.11, 5.12 e 5.13, foram obtidos pedindo ao Recommender que recomendasse ontologias analisando os conjuntos de termos de cada agrupamento como “palavras-chave”.

O primeiro caso a ser analisado será aquele em que os termos foram analisados como “palavras-chave” e recomendada uma ontologia.

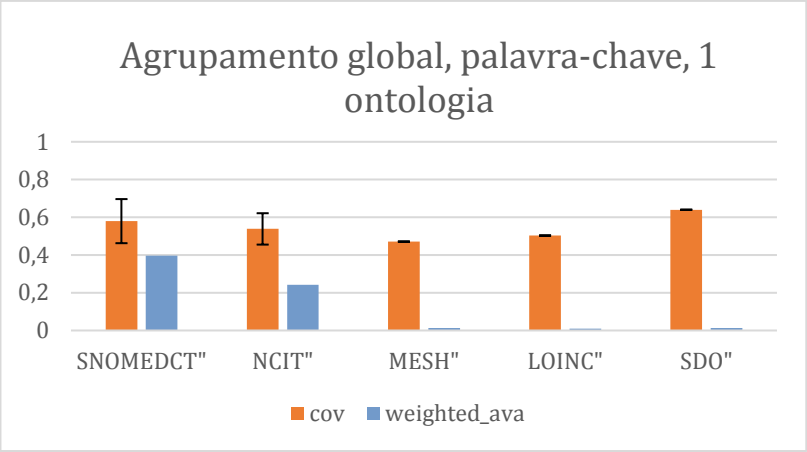

Figura 5.11- Gráfico de recomendações para o agrupamento global, palavra-chave, 1 ontologia

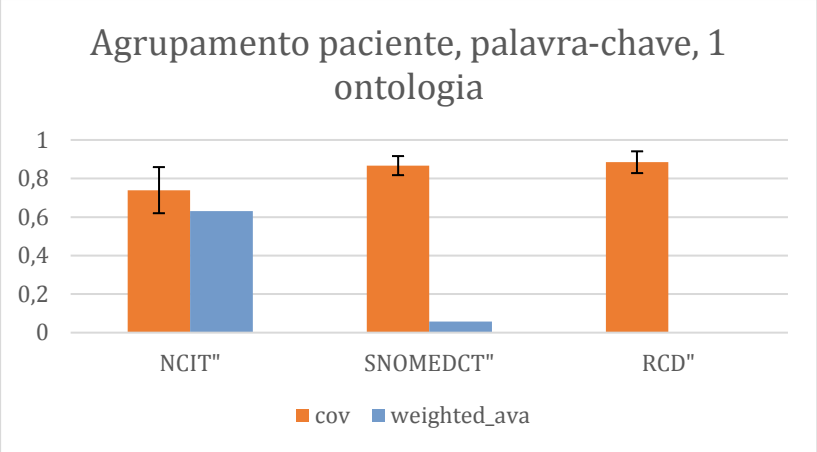

Figura 5.12- Gráfico de recomendações para o agrupamento paciente, palavra-chave, 1 ontologia

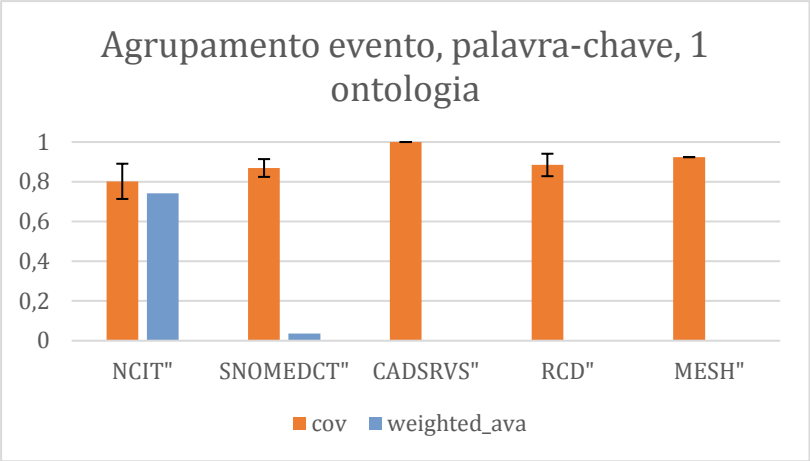

Figura 5.13- Gráfico de recomendações para o agrupamento evento, palavra-chave, 1 ontologia

Antes de analisar é importante relembrar que anotação por "palavras-chave" é diferente, e que essa diferença está na forma como um conceito é associado a um termo, um conceito só é associado a um termo se estes forem iguais na totalidade, isto vai influenciar o valor de cobertura médio. Visto que as anotações resultantes, apesar de serem em menor quantidade, no geral, terão uma pontuação mais alta do que anotações feitas com termos analisados como "texto".

Comparando este caso, com o caso "texto" onde se pedia para recomendar uma ontologia, os valores de cobertura médios no caso "palavras-chaves" para os seus agrupamentos são quase todos superiores aos valores médios do caso "texto", menos para o agrupamento "global", quem tem um valor médio de 0.57 e no caso "texto", o mesmo agrupamento, tem uma pontuação de 0.68.

As ontologias são ligeiramente diferentes entre os agrupamentos, o agrupamento "paciente" e "evento" têm a ontologia NCIT em primeiro lugar, enquanto o "global" tem a ontologia SNOMED-CT, contudo são todas ontologias genéricas.

As diferentes pontuações entre "texto" e "palavras-chave" devem-se às anotações que são obtidas, as anotações do tipo "palavras-chave" têm pontuações muito mais altas que as anotações de tipo "texto".

O valor cobertura médio é diferente entre os diferentes agrupamentos e tem a mesma causa que nas recomendações do tipo "texto", está relacionado com o domínio a que pertencem os termos e como é que estes estão relacionados entre si.

As próximas recomendações a serem analisadas e representadas pelas Figuras 5.14, 5.15 e 5.16, são do tipo "palavras-chave" e 2 ontologias.

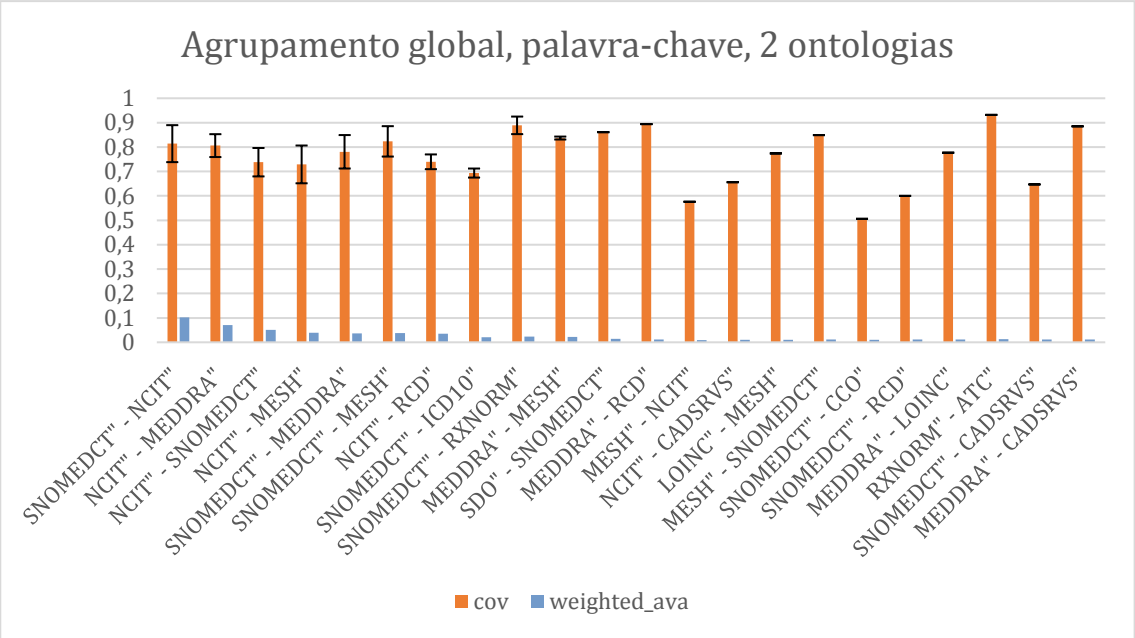

Figura 5.14 - Gráfico de recomendações para o agrupamento global, palavra-chave, 2 ontologias

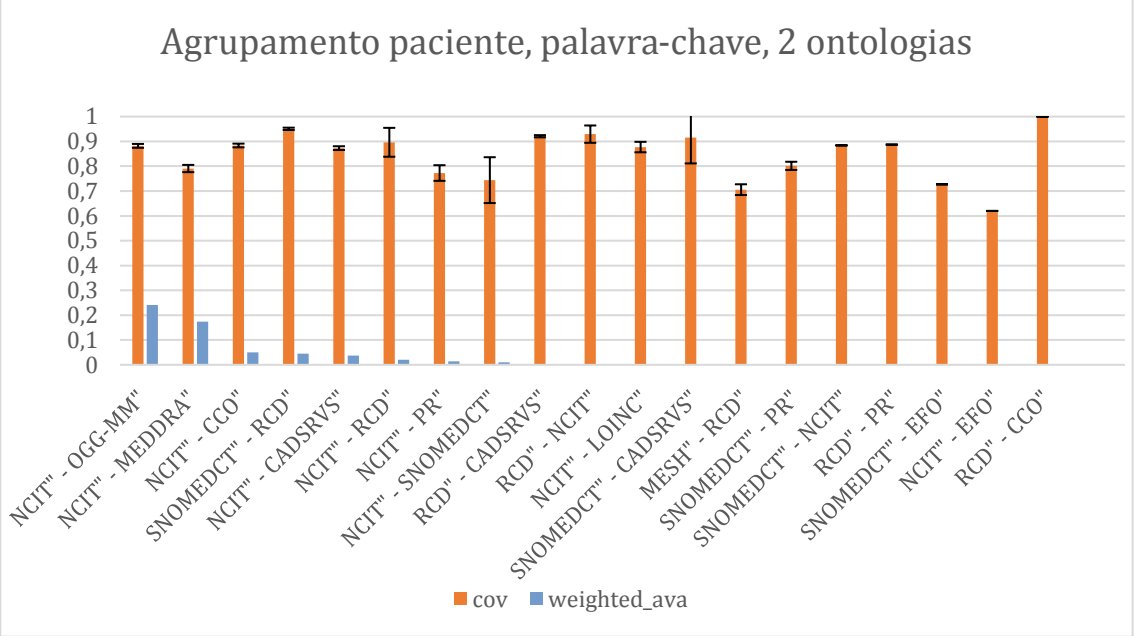

Figura 5.15 - Gráfico de recomendações para o agrupamento paciente, palavra-chave, 2 ontologias

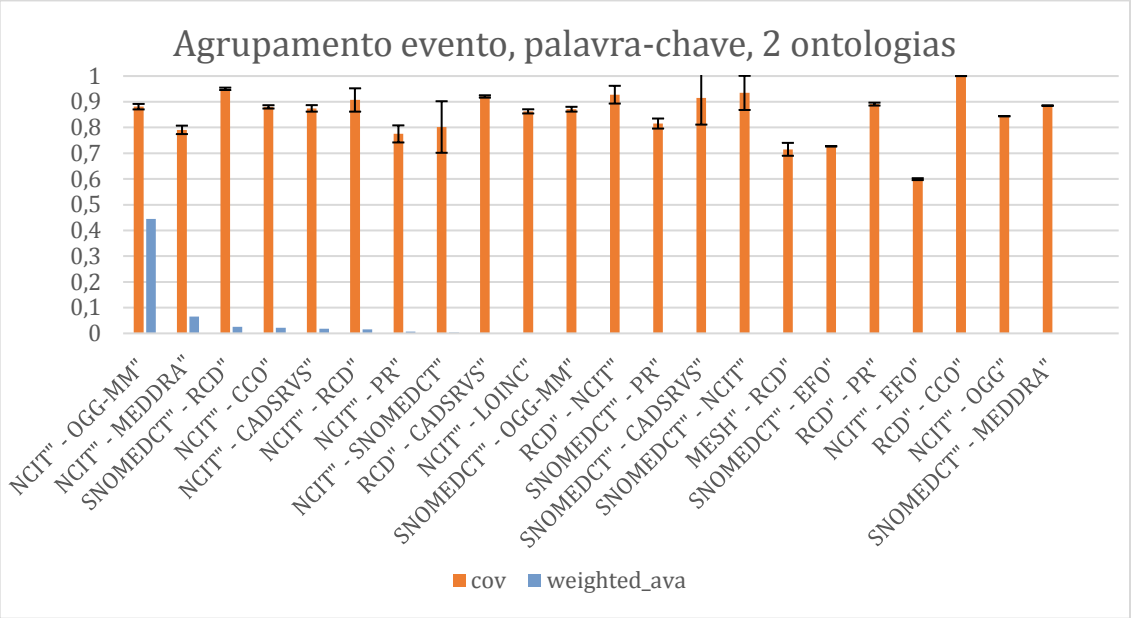

Figura 5.16 - Gráfico de recomendações para o agrupamento eventos, palavra-chave, 2 ontologias

Como é possível ver pelas tabelas, o valor médio aumentou com o uso de mais uma ontologia, e é de realçar que, mais uma vez, o agrupamento "global" tem em primeiro lugar duas ontologias genéricas, a NCIT e SNOMED-CT e os outros agrupamentos têm uma ontologia genérica a NCIT e uma específica, a OGG-MM.

Comparando os diferentes tipos de análise, o valor médio para as ontologias que ficaram em primeiro lugar no tipo "palavras-chave" e duas ontologias, foram 0.81, 0.88, 0.88 para o "global", "paciente" e "evento", respectivamente. E para o tipo "texto" foram 0.79, 0.81, 0.84, para o "global", "paciente" e "evento", respectivamente. Estes valores indicam que a anotação por "palavras-chaves" resulta em melhores anotações.

O próximo caso a ser analisado e representado pelas Figuras 5.17, 5.18 e 5.19, são as recomendações para 3 ontologias em que os termos foram analisados como "palavras-chave".

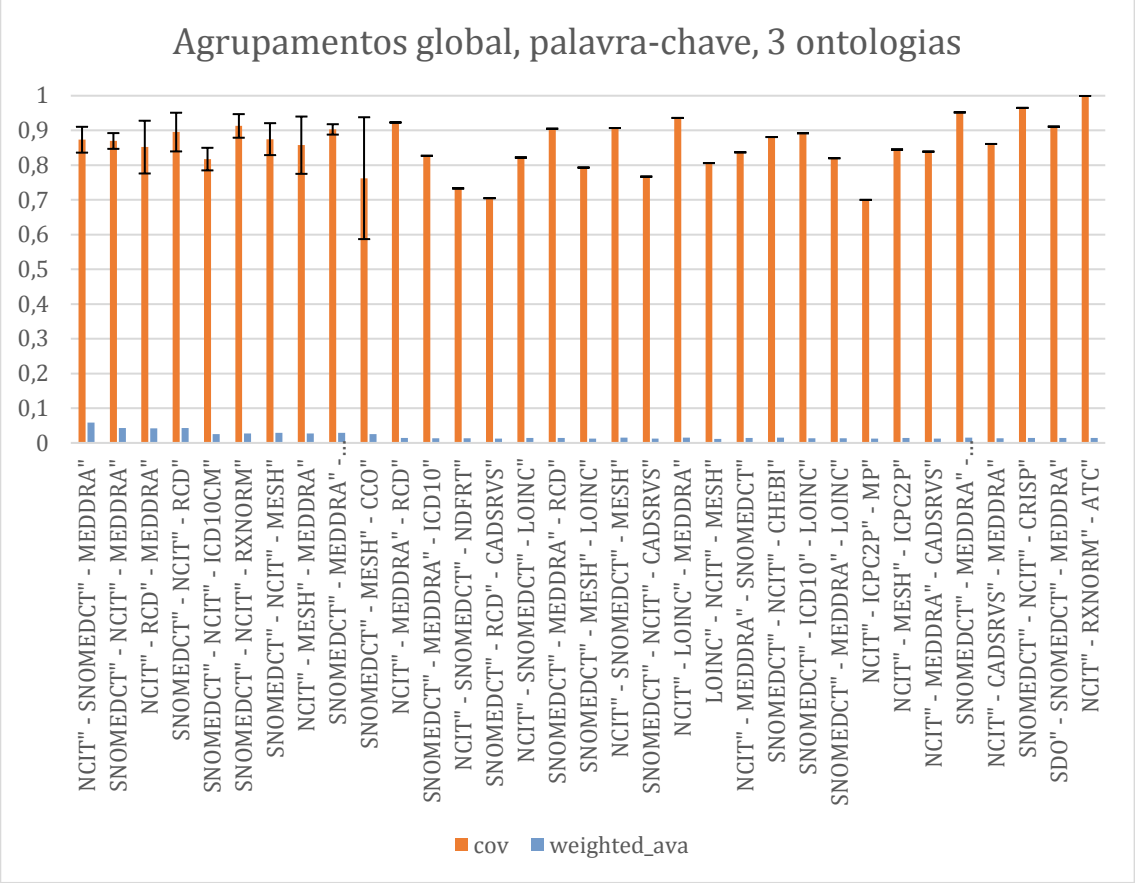

Figura 5.17 - Gráfico de recomendações para o agrupamento global, palavra-chave, 3 ontologias

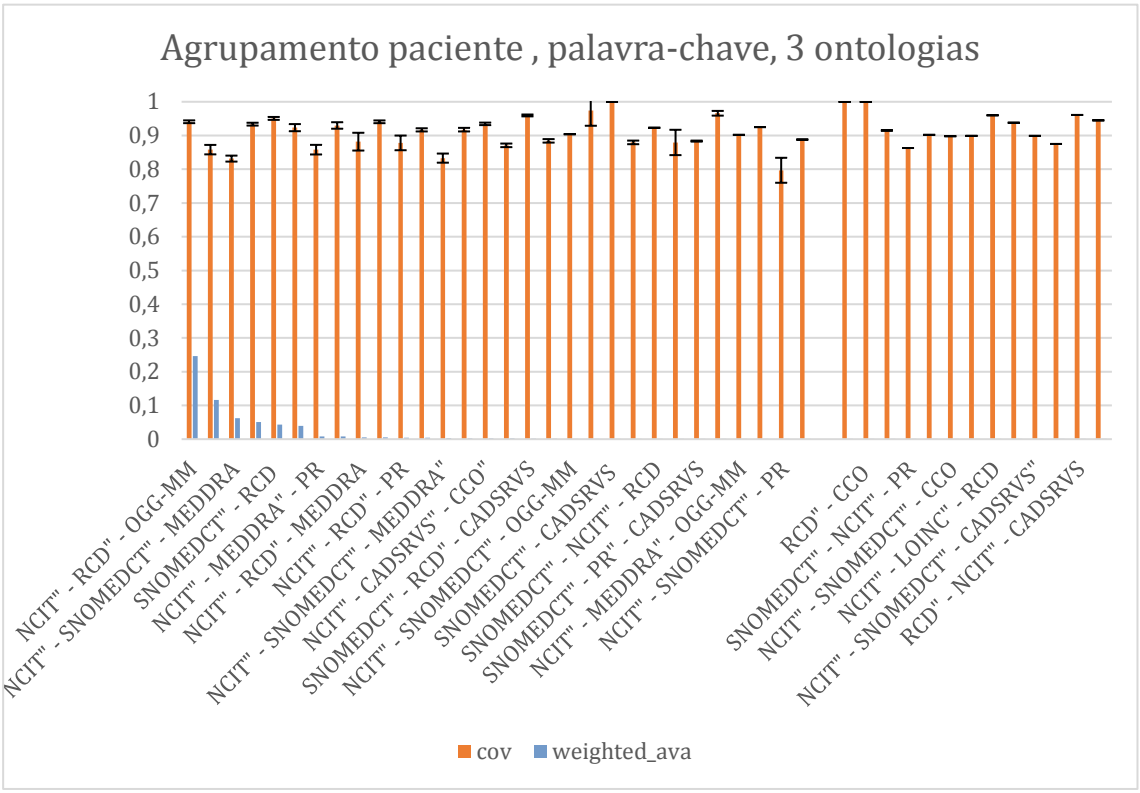

Figura 5.18 - Gráfico de recomendações para o agrupamento paciente, palavra-chave, 3 ontologias

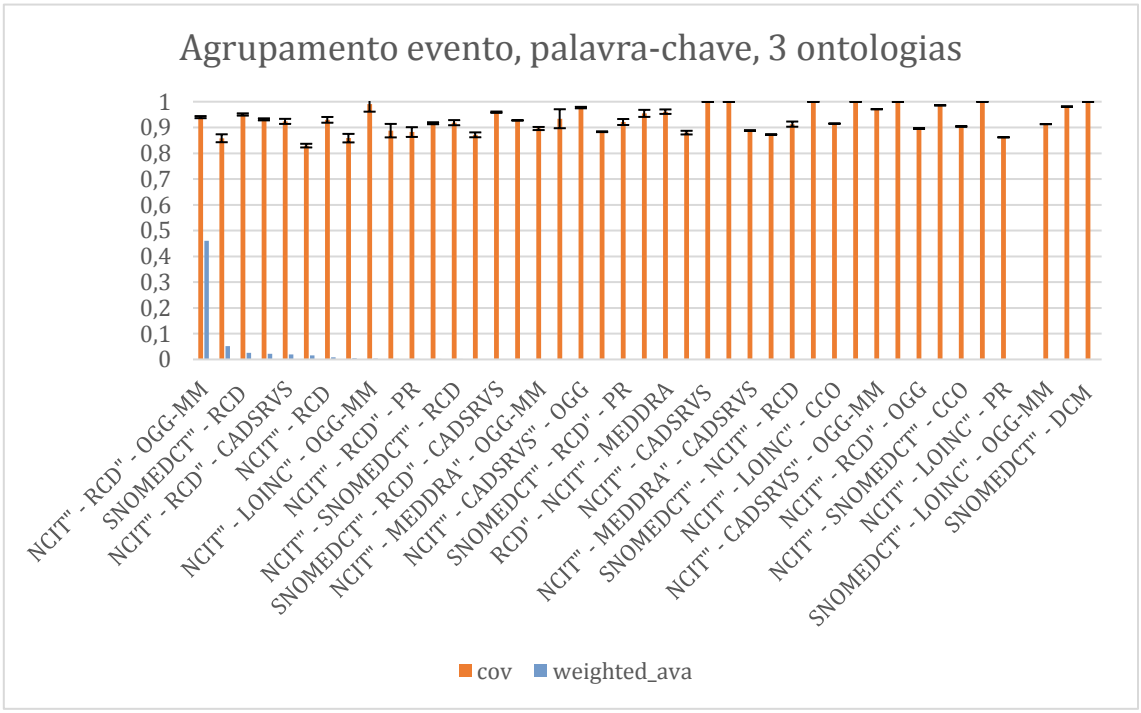

Figura 5.19 - Gráfico de recomendações para o agrupamento evento, palavra-chave, 3 ontologias

Mais uma vez o valor médio volta a aumentar com o crescimento de uma ontologia. O "global" tem como ontologias em primeiro lugar a NCIT, SNOMED-CT e MEDDRA, o valor médio de cobertura é de 0.87, o agrupamento "paciente" tem a NCIT, RCD e a OGG-MM, o valor médio de cobertura é de 0.94, por último o agrupamento "evento" teve em primeiro lugar também a NCIT, RCD e a OGG-MM, e o valor médio cobertura foi de 0.94.

As ontologias recomendadas para o "global" são quase todas genéricas, tirando a MEDDRA, que é uma ontologia com termos para regular actividades médicas, os outros agrupamentos têm a NCIT, que é uma ontologia genérica, a RCD é uma ontologia com termos clínicos e a OGG-MM é uma ontologia com genes e genomas de organismos.

Comparando os dois tipos de análise de termos, "texto" e "palavras-chave", para o caso em que foram recomendadas 3 ontologias, o valor médio dos conjuntos de ontologias que estavam em primeiro lugar para o tipo "texto" foram de 0.87, 0.89 e 0.90, para o "global", "paciente" e "evento", respectivamente.

Desta análise descobriu-se que a melhor forma de analisar e a que resulta em melhores anotações, é a de tipo "palavras-chave" e outra forma de melhorar o valor cobertura médio e que funciona em ambos os tipos de análise, é aumentar o número de ontologias usadas. Quanto maior for o número de ontologias usadas maior será o valor de cobertura (valor de cobertura do Recommender).

Com a análise destes dados não é possível concluir qual dos agrupamentos é o melhor, visto que os valores de cobertura médios obtidos foram muito semelhantes entre cada um dos agrupamentos.

5.4 Resultados do EHRannotator – casos de estudo

Como foi explicado no capítulo de desenho e no capítulo implementação, o EHRannotator utiliza as ontologias que foram mais vezes recomendadas para uma determinada situação para fazer a anotação dos dados.

Neste subcapítulo serão analisados e discutidos alguns testes que foram realizados (as anotações e resultados encontram-se nos anexos A e B), a nível dos agrupamentos paciente e evento. Importa ressalvar, que os dados utilizados não são reais, pelo que as conclusões a seguir discutidas apenas exemplificam o tipo de informação que é possível extrair usando a metodologia proposta neste trabalho.

Os testes têm como foco dois pacientes, o paciente com id 2 e o paciente com o id 17, e os seus "eventos", o evento com id 26999 para o paciente 2 e o evento com id 8779 para o paciente 17.

Os termos nestes agrupamentos foram analisados como "texto" e "palavras-chave", anotados com uma ontologia e três ontologias, o critério no processo de filtragem das anotações foi "especificidade".

Na Tabela 5.1 encontram-se todos os valores de cobertura obtidos para cada um dos testes no EHRannotator e na Tabela 5.2 encontram-se os valores de cobertura médios obtidos pelo RecommendOntologies para os diferentes agrupamentos, número de ontologias e técnicas de anotação.

Análise / Teste	"texto" 1 ontologia	"texto" 3 ontologias	"palavras-chave" 1 ontologia	"palavras-chave" 3 ontologias
Paciente 2	0,704	0,711	0,228	0,228
Paciente 17	0,712	0,722	0,209	0,209
Evento 26999	0,69	0,701	0,236	0,236
Evento 8779	0,717	0,727	0,214	0,214

Tabela 5.1- valores de cobertura obtidos no EHRannotator

Análise / Agrupamento	"texto" 1 ontologia	"texto" 3 ontologias	"palavras-chave" 1 ontologia	"palavras-chave" 3 ontologias
Paciente	0,73	0,89	0,73	0,94
Evento	0,74	0,9	0,8	0,94

Tabela 5.2 valores de cobertura médio obtidos pelo RecommendOntologies

É de salientar que os valores de cobertura são calculados através de formas diferentes; o valor de cobertura da Tabela 5.2 informa a razão da qualidade das anotações face à qualidade das anotações usando todas as ontologias, enquanto o valor da Tabela 5.1 informa a razão entre os termos cobertos com conceitos e os termos não cobertos com conceitos.

Comparando os resultados obtidos no EHRannotator com os resultados obtidos pelo RecommendOntologies chega-se a conclusões contraditórias quanto à melhor forma de analisar os dados. Segundo os dados do RecommendOntologies a melhor técnica para analisar os termos é "palavras-chave" porque apresenta valores de cobertura médios superiores aos valores obtidos analisando os termos com a técnica "texto". Contudo, essa conclusão verifica-se inválida nos testes feitos com o

EHRannotator, pois os casos em que os termos foram analisados com a técnica "texto" obtiveram valores de cobertura superiores aos casos em que os termos foram analisados com a técnica "palavras-chave". Esta diferença acontece porque a anotação feita por "palavras-chave" não é tão flexível como na anotação por "texto", o que torna difícil encontrar um conceito dentro das ontologias que seja completamente igual a um termo dado como *input*, logo não existem anotações parciais, e isso também justifica a diminuição do número de anotações e a diminuição do valor de cobertura nos testes de "texto" para os testes de "palavras-chave".

Tentado perceber qual a melhor forma de agrupar os termos (contexto), e através da observação apenas da Tabela 5.2, percebe-se que é difícil escolher qual é o melhor agrupamento pois a diferença entre os valores de cobertura obtidos para os dois agrupamentos é pouco relevante. Esta conclusão ganha mais solidez quando se analisa a Tabela 5.1, pois verifica-se que a diferença continua a ser mínima entre os dois agrupamentos. É importante salientar que não se pretende fazer uma generalização com os dados da Tabela 5.1, mas utilizá-los para verificar se os dados obtidos estão de acordo com as conclusões obtidas na Tabela 5.2. Como os dois contextos têm valores semelhantes tanto na Tabela 5.2 e na Tabela 5.1, pode-se concluir que as duas formas são adequadas para anotar a informação.

Uma conclusão errada que se tira por analisar a Tabela 5.2, é que a anotação por "palavra-chave" gera melhores e mais anotações e que por isso é a melhor técnica de anotar. Esta conclusão está errada, porque através da fórmula do Recommender obtém-se a qualidade das anotações obtidas em relação às melhores anotações obtidas usando todas as ontologias. Contudo, esta fórmula não tem em conta os termos que foram cobertos e os termos que não foram cobertos, sendo assim não se tem uma noção da percentagem de termos cobertos com conceitos. Usando a fórmula do EHRannotator o utilizador obtém o valor de cobertura da ontologia sobre os dados de *input* e com esse valor percebe a capacidade da ontologia para anotar o *input*.

E o que acontece para que os valores de cobertura médios calculados pelo Recommender serem diferentes para cada uma das técnicas deve-se ao facto da forma como estas técnicas combinam os conceitos com os termos e a facilidade que têm em gerar novas anotações. A técnica "palavra-chave", gera um menor número de anotações, mesmo com o acréscimo de novas ontologias, logo a pontuação total das anotações da ontologia recomendada é quase igual à pontuação obtida usando todas as ontologias, o que não acontece com a técnica "texto". Sendo assim o valor de cobertura médio para "palavra-chave" calculado pelo Recommender será quase sempre mais alto que o valor de cobertura médio obtido usando a técnica "texto". A técnica de "texto" tem maior liberdade para associar partes de termos a conceitos, logo cria mais anotações e para se

obter um valor de cobertura (calculado pelo Recommender) alto seria necessário utilizar mais ontologias e usar ontologias que possivelmente não estão no topo das recomendadas mas que têm conceitos associados aos termos de *input*, o que dificulta igualar a pontuação das anotações obtidas com a pontuação das anotações obtidas usando todas as ontologias.

Observando a Tabela 5.1 verifica-se que a técnica por "texto" cobre mais termos que a técnica por "palavra-chave", mas pela Tabela 5.2 percebe-se que a qualidade das anotações geradas pela análise "texto" é inferior à das anotações criadas com a técnica "palavra-chave". O que levanta uma questão: "será que as anotações feitas com a técnica "palavra-chave" têm menos qualidade que as anotações geradas usando a técnica "texto"?"

Esta questão é facilmente respondida cruzando as anotações (ver anexo A) que foram geradas através de "texto" e as que foram obtidas por "palavras-chave" para o mesmo caso. Por exemplo, para um evento usado nos testes verifica-se que as anotações geradas por "palavras-chave" estão presentes nas anotações geradas por "texto". Logo o conjunto das anotações geradas por "palavras-chave" está contido no conjunto de anotações geradas por "texto". Sendo assim a melhor técnica para analisar os termos é através de "texto", pois gera anotações com a mesma qualidade que a técnica "palavra-chave" e também gera mais anotações.

Capítulo 6

Conclusão

O propósito desta tese era desenvolver uma estratégia que anotasse os termos de vocabulários controlados utilizados nos registos electrónicos de saúde com conceitos de ontologias. Esta estratégia teria que dar resposta às seguintes questões: "Que agrupamento/contexto utilizar?", "Que ontologias escolher para fazer a anotação?", "Que técnicas de anotação usar?", "Qual o melhor conceito para um termo?", "Como combinar as técnicas de anotação com o processo de selecção da melhor ou melhores ontologias?".

A primeira questão fica sem resposta porque através dos testes feitos usando os dados do Openmrs, talvez por serem simulados, ambos os contextos ("evento" e "paciente") obtiveram valores de cobertura e valores de cobertura média muito semelhantes entre si em todos os testes. No entanto, a utilização da metodologia proposta utilizando dados reais permitirá saber qual a melhor escolha de agrupamento.

Para a selecção de ontologias, o sistema consegue recomendar ontologias ou conjuntos de ontologias para qualquer contexto baseando-se em 3 contextos diferentes: "global", "paciente" e "evento". Um cenário diferente daquele que foi utilizado no EHRannotator, e que para o qual esta estratégia teria os mesmo resultado, seria um cenário seria em que a base de dados não continha pacientes nem eventos, logo as recomendações seriam obtidas através de uma análise apenas ao vocabulário controlado.

Como foi explicado no capítulo 4 e verificado no capítulo 5, o valor de cobertura do Recommender reflecte a razão entre a qualidade das anotações obtidas usando as ontologias escolhidas e a pontuação total usando todas as ontologias, logo esta fórmula não reflecte a forma como essas anotações cobrem os termos de *input*. Esse facto é facilmente constatado percebendo a fórmula usada pelo Recommender e comparando, por exemplo, os valores de cobertura médios obtidos no RecommendOntologies para o caso "evento", "palavras-chave" e uma ontologia com os valores de cobertura obtidos nos testes do EHRannotator "evento", "palavras-chave" e uma ontologia. Após a recolha dos dados fornecidos pelo RecommendOntologies a primeira conclusão é que a

melhor forma de analisar os dados seria usando a técnica "palavras-chave", pois esta técnica de analisar os termos tem um valor de cobertura médio superior ao valor de cobertura médio obtido usando a técnica "texto". Com os resultados obtidos no EHRannotator verificamos que esta conclusão é na verdade inválida, pois o número de termos efectivamente cobertos através da análise por "palavra-chave" é muito inferior ao número de termos cobertos usando a técnica "texto" e além disso as anotações resultantes da análise por "palavras-chave" são um subconjunto do conjunto de anotações resultantes da análise por "texto", o que significa que a melhor forma de analisar os termos é através de "texto".

A solução para resolver o problema de um termo ter mais que um conceito associado passou por adicionar a cada conceito 3 valores diferentes, valor de centralidade, valor de especificidade e valor de densidade. Através de um processo de filtragem seleciona-se o conceito com maior valor segundo um determinado critério, escolhido pelo utilizador. Selecionar um critério é uma questão subjectiva, não existe um critério melhor que outro para filtrar os conceitos, isso depende dos objectivos do utilizador. E como estes critérios resolvem o problema de selecionar conceitos para os termos, estes critérios destacam esta estratégia de anotação de outras estratégias de anotação.

Existe um problema ligado com o uso da técnica de anotação "texto". Ao usar-se esta técnica serão geradas muitas anotações parciais, ou seja, cada palavra do termo é ligada a um conceito e isso não acontece usando a técnica de anotação "palavra-chave", pois esta só gera anotações completas. Este problema, não deve ser visto como algo negativo, mas sim como ideia para trabalho futuro. Uma possível tentativa de solução seria usar os casos em que existem termos com conceitos ligados a cada palavra do termo e substituir todos os conceitos por um único conceito que fosse semanticamente parecido com o termo de *input*.

Como foi demonstrado, esta estratégia pode adaptar as recomendações a qualquer contexto ("paciente", "evento") ou a nenhum contexto, ou seja, o sistema pode simplesmente analisar os termos existentes no vocabulário controlado e recomendar as ontologias, o que torna o processo de recomendação mais rápido. É inovadora porque não existe nenhum sistema que faça a filtragem das anotações usando os critérios de centralidade, densidade e especificidade. Permite a análise dos dados através de diferentes técnicas e como foi constatado, é necessário melhorar a técnica de anotação de "texto" de maneira que o sistema transforme as anotações parciais em anotações completas. Em suma, é uma estratégia que apresenta uma grande utilidade para sistemas que necessitem de dados semanticamente anotados para a extracção de conhecimentos, como por exemplo, sistemas de suporte à decisão. Como é baseada na exploração de

recursos externos existentes, não é necessário manter versões locais das ontologias biomédicas, com as vantagens em termos de actualização e armazenamento que isso implica.

Capítulo 7

Bibliografia

[1] Razzaque, A., & Karolak, M. (2011). Knowledge management and electronic health record facilitate clinical support to improve healthcare quality. In International Conference on E-business, Management and Economics IPEDR (Vol. 3, pp. 238-242).

[2] Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., & Taylor, R. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Affairs, 24(5), pp. 1103-1117.

[3] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), pp. 395-405.

[4] Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I., & Novak, P. K. (2011, January). Using ontologies in semantic data mining with segs and g-segs. InDiscovery Science (pp. 165-178). Springer Berlin Heidelberg.

[5] Holmes, A. B., Hawson, A., Liu, F., Friedman, C., Khiabanian, H., & Rabadan, R. (2011). Discovering disease associations by integrating electronic clinical data and medical literature. PLoS One, 6(6), e21132.

[6] Häyrinen, K., Saranto, K., & Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the research literature. International journal of medical informatics, 77(5), pp. 291-304.

[7] Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. The knowledge engineering review, 11(02), pp. 93-136.

[8] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semantics: science, services and agents on the World Wide Web, 4(1), pp. 14-28

[9] Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M. A., & Musen, M. (2009). NCBO annotator: semantic annotation of biomedical data. InInternational Semantic Web Conference, Poster and Demo session (Vol. 110).

[10] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), pp. 507-513.

[11] Patel, C., Supekar, K., Lee, Y., & Park, E. K. (2003, November). OntoKhoj: a semantic web portal for ontology searching, ranking and classification. InProceedings of the 5th ACM international workshop on Web information and data management (pp. 58-61). ACM.

[12] Alani, H., & Brewster, C. (2006). Metrics for ranking ontologies.

[13] Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39(suppl 2), W541-W545.

[14] Jonquet, C., Shah, N. H., & Musen, M. A. (2009). Prototyping a biomedical ontology recommender service. Bio-Ontologies: Knowledge in Biology, ISMB/ECCB SIG.

[15] Butt, A. S., Haller, A., & Xie, L. (2015). DWRank: Learning concept ranking for ontology search. Semantic Web, (Preprint), pp. 1-15.

[16] Peroni, S., Motta, E., & d'Aquin, M. (2008, December). Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. In Asian Semantic Web Conference (pp. 242-256). Springer Berlin Heidelberg.

[17] Huang, J., Huan, J., Tropsha, A., Dang, J., Zhang, H., & Xiong, M. (2013, December). Semantics-driven frequent data pattern mining on electronic health records for effective adverse drug event monitoring. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on (pp. 608-611). IEEE.

[18] Trifiro, G., Pariente, A., Coloma, P. M., Kors, J. A., Polimeni, G., Miremont-Salamé, G., & Caputi, A. P. (2009). Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor?.Pharmacoepidemiology and Drug Safety, 18(12), pp. 1176-1184.

[19] Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. S., Chen, P. J., & Huff, S. M. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association, 20(e2), e341-e348

[20] Jonquet, C., Musen, M. A., & Shah, N. H. (2010). Building a biomedical ontology recommender web service. Journal of biomedical semantics, 1(1), 1.

[21] Martínez-Romero, M., Vázquez-Naya, J. M., Pereira, J., & Pazos, A. (2012). A multi-criteria approach for automatic ontology recommendation using collective knowledge. In Recommender Systems for the Social Web (pp. 89-103). Springer Berlin Heidelberg.

[22] McBride, B. (2001, May). Jena: Implementing the RDF model and syntax specification. In Proceedings of the Second International Conference on Semantic Web-Volume 40 (pp. 23-28). CEUR-WS. org.

[23] Wolfe, B. A., Mamlin, B. W., Biondich, P. G., Fraser, H. S., Jazayeri, D., Allen, C., & Tierney, W. M. (2006). The OpenMRS system: collaborating toward an open source EMR for developing countries. In AMIA Annual Symposium Proceedings (Vol. 2006, p. 1146). American Medical Informatics Association.

Capítulo 8 Anexos

8.1 Anexo A

Neste anexo estão as anotações obtidas para cada teste feito com o EHRannotator. As tabelas contêm os dados sobre cada teste como id paciente/evento, técnica de anotação usada, número de ontologias, número de termos usados, número de anotações antes e depois da filtragem e cobertura.

Tabela 8.1- Teste paciente 2, 1 ontologia, técnica "texto"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
2	1	"texto"	57	0,704	122	71
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C25211	NCIT	0	0.25	0	0.25
Visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
New	C94522	NCIT	0	0.4375	0	0.4375
interval	C25543	NCIT	0.9334	0.25	0.0022	0.25
use	C95018	NCIT	0	0.1875	0	0.1875
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C29844	NCIT	0	0.3125	0	0.3125
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C29846	NCIT	0	0.3125	0.0004	0.3125
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875

dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	C23658	NCIT	0	0.1875	0	0.1875
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	C25208	NCIT	1	0.1875	0.0016	0.1875
current	C25471	NCIT	0	0.25	0	0.25
hiv	C14219	NCIT	0	0.4375	0.0007	0.4375
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient	C53691	NCIT	0	0.1875	0	0.1875
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C15843	NCIT	0.6102	0.25	0.0053	0.25
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	C25619	NCIT	1	0.1875	0.0062	0.1875
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
pregnancy	C25742	NCIT	0	0.25	0.0004	0.25
status	C25688	NCIT	0.903	0.125	0.0149	0.125
conception	C16465	NCIT	0	0.25	0	0.25
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd4	C103810	NCIT	0	0.4375	0	0.4375
count	C48485	NCIT	0	0.3125	0	0.3125
cd3	C103809	NCIT	0	0.4375	0	0.4375
lymph	C13252	NCIT	0	0.1875	0	0.1875

percent	C48570	NCIT	0	0.3125	0.0053	0.3125
cd8	C103811	NCIT	0	0.4375	0	0.4375
tests	C25294	NCIT	0.5725	0.25	0.0013	0.25
problem	C28020	NCIT	0.8334	0.125	0.0019	0.125
added	C45330	NCIT	0	0.25	0	0.25
diagnosis	C49653	NCIT	0	0.4375	0.0004	0.4375

Tabela 8.2- Teste paciente 17, 1 ontologia, técnica "texto"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
17	1	"texto"	43	0,712	99	56
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C25211	NCIT	0	0.25	0	0.25
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C25543	NCIT	0.9334	0.25	0.0022	0.25
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C29844	NCIT	0	0.3125	0	0.3125
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C29846	NCIT	0	0.3125	0.0004	0.3125
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	C23658	NCIT	0	0.1875	0	0.1875
rate	C77538	NCIT	0	0.1875	0	0.1875

systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	C25208	NCIT	1	0.1875	0.0016	0.1875
current	C25471	NCIT	0	0.25	0	0.25
hiv	C14219	NCIT	0	0.4375	0.0007	0.4375
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient	C53691	NCIT	0	0.1875	0	0.1875
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C15843	NCIT	0.6102	0.25	0.0053	0.25
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	C25619	NCIT	1	0.1875	0.0062	0.1875
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875

Tabela 8.3- Teste paciente 2, 3 ontologias, técnica "texto"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num de anotações filtradas
2	3	"texto"	57	0,711	233	75
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C0205539	HL7	0	0.5	0	0.5
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C1552713	HL7	0.7223	0.5	0.0119	0.5
use	C95018	NCIT	0	0.1875	0	0.1875
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C1561540	HL7	0	0.6	0	0.6

overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C1561542	HL7	0	0.6	0	0.6
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen saturation	LP21258-6	LOINC	0	0.0625	0	0.0625
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	8462-4	LOINC	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	LP7289-4	LOINC	0	0.375	0.0012	0.375
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	8480-6	LOINC	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	29463-7	LOINC	0	0.25	0	0.25
current	C25471	NCIT	0	0.25	0	0.25
who hiv stage	45233-4	LOINC	0	0.25	0	0.25
hiv	C0086413	HL7	0	0.5	0	0.5
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
drugs	LP18046-0	LOINC	0.0359	0.1875	0.003	0.1875
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient reported	C0747307	HL7	0	0.7	0	0.7
patient	C0030705	HL7	0	0.7	0	0.7
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C3853787	HL7	0	0.4	0	0.4
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	LP175754-3	LOINC	1	0.25	0.0003	0.25

treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
pregnancy	67471-3	LOINC	0	0.25	0	0.25
status	C25688	NCIT	0.903	0.125	0.0149	0.125
conception	C16465	NCIT	0	0.25	0	0.25
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd4	C103810	NCIT	0	0.4375	0	0.4375
count	C0750480	HL7	0	0.4	0	0.4
cd3	C103809	NCIT	0	0.4375	0	0.4375
lymph	C13252	NCIT	0	0.1875	0	0.1875
percent	C48570	NCIT	0	0.3125	0.0053	0.3125
cd8	C103811	NCIT	0	0.4375	0	0.4375
tests	C25294	NCIT	0.5725	0.25	0.0013	0.25
problem	75326-9	LOINC	0	0.25	0	0.25
added	C45330	NCIT	0	0.25	0	0.25
diagnosis	C0011900	HL7	0	0.6	0	0.6

Tabela 8.4- Teste paciente 17, 3 ontologias, técnica "texto"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
17	3	"texto"	43	0,722	181	60
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C0205539	HL7	0	0.5	0	0.5
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C1552713	HL7	0.7223	0.5	0.0119	0.5
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C1561540	HL7	0	0.6	0	0.6
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875

month	C1561542	HL7	0	0.6	0	0.6
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen saturation	LP21258-6	LOINC	0	0.0625	0	0.0625
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	8462-4	LOINC	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	LP7289-4	LOINC	0	0.375	0.0012	0.375
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	8480-6	LOINC	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	29463-7	LOINC	0	0.25	0	0.25
current	C25471	NCIT	0	0.25	0	0.25
who hiv stage	45233-4	LOINC	0	0.25	0	0.25
hiv	C0086413	HL7	0	0.5	0	0.5
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
drugs	LP18046-0	LOINC	0.0359	0.1875	0.003	0.1875
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient reported	C0747307	HL7	0	0.7	0	0.7
patient	C0030705	HL7	0	0.7	0	0.7
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C3853787	HL7	0	0.4	0	0.4
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	LP175754-3	LOINC	1	0.25	0.0003	0.25
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875

Tabela 8.5- Teste evento 26999, 1 ontologia, técnica "texto"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
26999	1	"texto"	55	0,69	118	68
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C25211	NCIT	0	0.25	0	0.25
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
pregnancy	C25742	NCIT	0	0.25	0.0004	0.25
status	C25688	NCIT	0.903	0.125	0.0149	0.125
conception	C16465	NCIT	0	0.25	0	0.25
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C25543	NCIT	0.9334	0.25	0.0022	0.25
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C29844	NCIT	0	0.3125	0	0.3125
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C29846	NCIT	0	0.3125	0.0004	0.3125
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	C23658	NCIT	0	0.1875	0	0.1875
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25

sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	C25208	NCIT	1	0.1875	0.0016	0.1875
current	C25471	NCIT	0	0.25	0	0.25
hiv	C14219	NCIT	0	0.4375	0.0007	0.4375
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd4	C103810	NCIT	0	0.4375	0	0.4375
count	C48485	NCIT	0	0.3125	0	0.3125
cd3	C103809	NCIT	0	0.4375	0	0.4375
lymph	C13252	NCIT	0	0.1875	0	0.1875
percent	C48570	NCIT	0	0.3125	0.0053	0.3125
cd8	C103811	NCIT	0	0.4375	0	0.4375
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient	C53691	NCIT	0	0.1875	0	0.1875
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C15843	NCIT	0.6102	0.25	0.0053	0.25
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	C25619	NCIT	1	0.1875	0.0062	0.1875
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
tests	C25294	NCIT	0.5725	0.25	0.0013	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875

Tabela 8.6- Teste evento 8779, 1 ontologia, técnica "texto"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
8779	1	"texto"	42	0,717	97	55
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C25211	NCIT	0	0.25	0	0.25
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875

new	C94522	NCIT	0	0.4375	0	0.4375
interval	C25543	NCIT	0.9334	0.25	0.0022	0.25
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C29844	NCIT	0	0.3125	0	0.3125
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C29846	NCIT	0	0.3125	0.0004	0.3125
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	C23658	NCIT	0	0.1875	0	0.1875
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	C25208	NCIT	1	0.1875	0.0016	0.1875
current	C25471	NCIT	0	0.25	0	0.25
hiv	C14219	NCIT	0	0.4375	0.0007	0.4375
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient	C53691	NCIT	0	0.1875	0	0.1875
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25

cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C15843	NCIT	0.6102	0.25	0.0053	0.25
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	C25619	NCIT	1	0.1875	0.0062	0.1875
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25

Tabela 8.7- Teste evento 26999, 3 ontologias, técnica "texto"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
26999	3	"texto"	55	0,701	221	72
	Anotações filtradas					
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C0205539	HL7	0	0.5	0	0.5
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
pregnancy	67471-3	LOINC	0	0.25	0	0.25
status	C25688	NCIT	0.903	0.125	0.0149	0.125
conception	C16465	NCIT	0	0.25	0	0.25
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C1552713	HL7	0.7223	0.5	0.0119	0.5
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C1561540	HL7	0	0.6	0	0.6
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C1561542	HL7	0	0.6	0	0.6
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen saturation	LP21258-6	LOINC	0	0.0625	0	0.0625
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	8462-4	LOINC	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875

dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	LP7289-4	LOINC	0	0.375	0.0012	0.375
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	8480-6	LOINC	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	29463-7	LOINC	0	0.25	0	0.25
current	C25471	NCIT	0	0.25	0	0.25
who hiv stage	45233-4	LOINC	0	0.25	0	0.25
hiv	C0086413	HL7	0	0.5	0	0.5
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd4	C103810	NCIT	0	0.4375	0	0.4375
count	C0750480	HL7	0	0.4	0	0.4
cd3	C103809	NCIT	0	0.4375	0	0.4375
lymph	C13252	NCIT	0	0.1875	0	0.1875
percent	C48570	NCIT	0	0.3125	0.0053	0.3125
cd8	C103811	NCIT	0	0.4375	0	0.4375
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
drugs	LP18046-0	LOINC	0.0359	0.1875	0.003	0.1875
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient reported	C0747307	HL7	0	0.7	0	0.7
patient	C0030705	HL7	0	0.7	0	0.7
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C3853787	HL7	0	0.4	0	0.4
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	LP175754-3	LOINC	1	0.25	0.0003	0.25
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25
tests	C25294	NCIT	0.5725	0.25	0.0013	0.25
ordered	C48906	NCIT	0	0.1875	0	0.1875

Tabela 8.8- Teste evento 8779, 3 ontologias, técnica "texto"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
8779	3	"texto"	42	0,727	179	59
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
scheduled	C0205539	HL7	0	0.5	0	0.5
visit	C39564	NCIT	0	0.3125	0	0.3125
category	C115468	NCIT	0	0.1875	0	0.1875
discordant	C101125	NCIT	0	0.125	0	0.125
couple	C61302	NCIT	0	0.1875	0	0.1875
new	C94522	NCIT	0	0.4375	0	0.4375
interval	C1552713	HL7	0.7223	0.5	0.0119	0.5
use	C95018	NCIT	0	0.1875	0	0.1875
primary	C25251	NCIT	0	0.25	0	0.25
regimen	C15697	NCIT	0	0.3125	0.0038	0.3125
adherence	C25729	NCIT	0	0.1875	0	0.1875
past week	C95402	NCIT	0	0.3125	0	0.3125
past	C25609	NCIT	0	0.25	0	0.25
week	C1561540	HL7	0	0.6	0	0.6
overall	C25605	NCIT	0	0.25	0	0.25
drug	C459	NCIT	1	0.1875	0.0007	0.1875
month	C1561542	HL7	0	0.6	0	0.6
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
blood	C12434	NCIT	0	0.1875	0	0.1875
oxygen saturation	LP21258-6	LOINC	0	0.0625	0	0.0625
oxygen	C722	NCIT	0	0.3125	0	0.3125
saturation	C61427	NCIT	0	0.1875	0	0.1875
diastolic blood pressure	8462-4	LOINC	0	0.25	0	0.25
diastolic	C62109	NCIT	0	0.25	0	0.25
blood pressure	C54707	NCIT	0.32	0.3125	0.0007	0.3125
pressure	C25195	NCIT	1	0.1875	0.0004	0.1875
dbp	C84254	NCIT	0	0.375	0	0.375
pulse	C49676	NCIT	0	0.4375	0	0.4375
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
heart	LP7289-4	LOINC	0	0.375	0.0012	0.375
rate	C77538	NCIT	0	0.1875	0	0.1875
systolic blood pressure	8480-6	LOINC	0	0.25	0	0.25
systolic	C62110	NCIT	0	0.25	0	0.25
sbp	C99624	NCIT	0	0.25	0	0.25
temperature	C25206	NCIT	1	0.1875	0.001	0.1875
temp	C10136	NCIT	0	0.25	0	0.25
weight	29463-7	LOINC	0	0.25	0	0.25

current	C25471	NCIT	0	0.25	0	0.25
who hiv stage	45233-4	LOINC	0	0.25	0	0.25
hiv	C0086413	HL7	0	0.5	0	0.5
stage	C16899	NCIT	0	0.25	0.001	0.25
cdc	C123782	NCIT	0	0.4375	0	0.4375
staging	C15608	NCIT	0.3334	0.25	0.001	0.25
criteria	C25466	NCIT	1	0.125	0.0044	0.125
met	C122238	NCIT	0	0.5	0	0.5
return	C71900	NCIT	0	0.1875	0	0.1875
visit date	C83031	NCIT	0	0.3125	0	0.3125
date	C72063	NCIT	0	0.375	0	0.375
drugs	LP18046-0	LOINC	0.0359	0.1875	0.003	0.1875
treatment	C49656	NCIT	0	0.4375	0.0004	0.4375
patient reported	C0747307	HL7	0	0.7	0	0.7
patient	C0030705	HL7	0	0.7	0	0.7
reported	C25375	NCIT	0.6452	0.25	0.0115	0.25
cryptococcus	C77184	NCIT	0	0.3125	0.0016	0.3125
pcp	C74694	NCIT	0	0.625	0	0.625
prophylaxis	C3853787	HL7	0	0.4	0	0.4
tuberculosis	C3423	NCIT	0	0.5625	0.0041	0.5625
plan	LP175754-3	LOINC	1	0.25	0.0003	0.25
treatment plan	C60735	NCIT	0	0.25	0.0004	0.25

Tabela 8.9- Teste paciente 2, 1 ontologia, técnica "palavra-chave"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
2	1	"palavra-chave"	57	0,236	13	13
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
dbp	C84254	NCIT	0	0.375	0	0.375
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
sbp	C52267	NCIT	0	0.0625	0	0.0625
systolic	C62110	NCIT	0	0.25	0	0.25
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375

cd4	C21362	NCIT	0	0.25	0	0.25
cd8	C51135	NCIT	0	0.0625	0	0.0625

Tabela 8.10- Teste paciente 17, 1 ontologia, técnica "palavra-chave"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
17	1	"palavra-chave"	43	0,209	9	9
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
dbp	C84254	NCIT	0	0.375	0	0.375
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
sbp	C52267	NCIT	0	0.0625	0	0.0625
systolic	C62110	NCIT	0	0.25	0	0.25

Tabela 8.11- Teste paciente 2, 3 ontologias, técnica "palavra-chave"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
2	3	"palavra-chave"	57	0,228	21	13
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd8	C51135	NCIT	0	0.0625	0	0.0625
blood oxygen saturation	X7708	RCD	0	0.4706	0.0371	0.4706
diastolic blood pressure	XM02Y	RCD	0	0.7059	0.0124	0.7059
systolic blood pressure	XM02X	RCD	0	0.7059	0.0124	0.7059
systolic	X78u9	RCD	0	0.4118	0.0926	0.4118
dbp	OGG_3000013170	OGG-MM	0	0.9	0	0.9
sbp	OGG_3000020234	OGG-MM	0	0.9	0	0.9

cd4	OGG_3000012504	OGG-MM	0	0.9	0	0.9

Tabela 8.12- Teste paciente 17, 3 ontologias, técnica "palavra-chave"

Paciente	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
17	3	"palavra-chave"	43	0,209	16	9
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
blood oxygen saturation	X7708	RCD	0	0.4706	0.0371	0.4706
diastolic blood pressure	XM02Y	RCD	0	0.7059	0.0124	0.7059
systolic blood pressure	XM02X	RCD	0	0.7059	0.0124	0.7059
systolic	X78u9	RCD	0	0.4118	0.0926	0.4118
dbp	OGG_3000013170	OGG-MM	0	0.9	0	0.9
sbp	OGG_3000020234	OGG-MM	0	0.9	0	0.9

Tabela 8.13- Teste evento 26998, 1 ontologia, técnica "palavra-chave"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
26999	1	"palavra-chave"	55	0,236	13	13
Anotações filtradas						
Termo	Conceito	Ontologia	Cen.	Esp.	Den.	Ava. final
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
dbp	C84254	NCIT	0	0.375	0	0.375
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
sbp	C52267	NCIT	0	0.0625	0	0.0625
systolic	C62110	NCIT	0	0.25	0	0.25
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375

cd4	C21362	NCIT	0	0.25	0	0.25
cd8	C51135	NCIT	0	0.0625	0	0.0625

Tabela 8.14 - Teste evento 8779, 1 ontologia, técnica "palavra-chave"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
8779	1	"palavra-chave"	42	0,214	9	9
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
blood oxygen saturation	C60832	NCIT	0	0.4375	0.0004	0.4375
diastolic blood pressure	C25299	NCIT	0	0.25	0	0.25
dbp	C84254	NCIT	0	0.375	0	0.375
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
systolic blood pressure	C25298	NCIT	0	0.25	0	0.25
sbp	C52267	NCIT	0	0.0625	0	0.0625
systolic	C62110	NCIT	0	0.25	0	0.25

Tabela 8.15- Teste evento 26999, 3 ontologias, técnica "palavra-chave"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
26999	3	"palavra-chave"	55	0,236	21	13
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
pregnancy status	C69218	NCIT	0	0.4375	0.0016	0.4375
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
cd4 count	C55931	NCIT	0	0.4375	0.0016	0.4375
cd8	C51135	NCIT	0	0.0625	0	0.0625
blood oxygen saturation	X7708	RCD	0	0.4706	0.0371	0.4706
diastolic blood pressure	XM02Y	RCD	0	0.7059	0.0124	0.7059
systolic blood pressure	XM02X	RCD	0	0.7059	0.0124	0.7059
systolic	X78u9	RCD	0	0.4118	0.0926	0.4118
dbp	OGG_3000013170	OGG-MM	0	0.9	0	0.9
sbp	OGG_3000020234	OGG-MM	0	0.9	0	0.9
cd4	OGG_3000012504	OGG-MM	0	0.9	0	0.9

Tabela 8.16- Teste evento 8779, 3 ontologias, técnica "palavra-chave"

Evento	Num. Ontologias	Técnica	Num. De Termos	Cobertura	Num. Anotações sem filtragem	Num. de anotações filtradas
8779	3	"palavra-chave"	42	0,214	16	9
Anotações filtradas						
Termo	Conceito	Ontologia	Centralidade	Esp.	Densidade	Ava. final
diastolic	C62109	NCIT	0	0.25	0	0.25
pulse	C64713	NCIT	0	0.1875	0	0.1875
heart rate	C49677	NCIT	0	0.4375	0.0019	0.4375
blood oxygen saturation	X7708	RCD	0	0.4706	0.0371	0.4706
diastolic blood pressure	XM02Y	RCD	0	0.7059	0.0124	0.7059
systolic blood pressure	XM02X	RCD	0	0.7059	0.0124	0.7059
systolic	X78u9	RCD	0	0.4118	0.0926	0.4118
dbp	OGG_3000013170	OGG-MM	0	0.9	0	0.9
sbp	OGG_3000020234	OGG-MM	0	0.9	0	0.9

8.2 Anexo B

Neste anexo estão os resultados dos testes feito com o EHRannotator, os resultados são constituídos pelo número de anotações não filtradas, número de anotações após filtragem, valor de cobertura e o número de termos usados como input. Em alguns dos testes são feitas observações e feitas explicações com o intuito de compreender os resultados obtidos.

O primeiro teste a ser analisado será aquele em que os termos foram analisados como "texto" e foi usada uma ontologia para fazer a anotação.

Observando os resultados do RecommenderOntologies, para este caso ("texto" e uma ontologia) a ontologia recomendada para os dois agrupamentos foi a NCIT, o valor médio de cobertura foi de 0.73 para o agrupamento "paciente" e 0.74 para o agrupamento "evento".

Os resultados da anotação dos termos do paciente 2, analisados como "texto" e anotados com NCIT foram os seguintes:

- Número de termos: 57;
- Número de anotações antes da filtragem: 122;
- Número de anotações após a filtragem: 71;
- Valor de cobertura: 0.704;

Como é possível ver pelos resultados do teste, o número de anotações é superior ao número de termos dados como *input* e isso devesse à forma como os termos são anotados, neste caso, existe a liberdade de associar um conceito a uma palavra dentro do termo e isto faz com o número de anotações final seja superior ao número de termos.O valor de cobertura é semelhante ao valor de cobertura médio do Recommender, mas é apenas uma coincidência, como se verá nos testes em que os termos são analisados como "palavras-chave".

Para o paciente 17, usando a mesma ontologia e a mesma análise de termos, os resultados foram os seguintes:

- Número de termos: 43;
- Número de anotações antes da filtragem: 99;
- Número de anotações após filtragem: 56;
- Valor de cobertura: 0.712;

Neste caso apesar de o número de termos de *input* serem iguais, todos os outros valores são diferentes, devido ao facto de os pacientes terem termos diferentes. O valor de cobertura tem uma diferença maior para o valor médio de cobertura calculado pelo Recommender (0.73) do que no paciente 2.

O próximo teste analisado será o paciente id 2, onde os termos foram analisados como "texto" e anotados com 3 ontologias, a NCIT, LOINC e HL7, os resultados foram os seguintes:

- Número de termos: 57;
- Número de anotações antes da filtragem: 233;
- Número de anotações após filtragem: 75;
- Valor de cobertura: 0.711;

Como é possível observar, e comparando estes dados com os dados anteriores do paciente 2 para "texto" e uma ontologia, ao usarem-se mais ontologias o número de anotações aumentou e o valor de cobertura também, passou de 0.704 para 0.711. Esse aumento reflete-se no número de anotações filtradas nos dois casos, passam de 71 para 75 anotações, isto significa que há termos que não são anotados pela NCIT mas pelas ontologias LOINC e a HL7.

Segundo os dados do Recommender, o valor médio de cobertura para o agrupamento "paciente" "texto" e 3 ontologias é de 0.89. A diferença entres estes dois valores é mais notória e isso deve-se à forma como os dois valores são calculados.

Os resultados do teste para o paciente 17 onde os termos foram analisados como "texto" e anotados com a NCIT, a LOINC e a HL7, foram:

- Número de termos: 43;
- Número de anotações antes da filtragem: 181;
- Número de anotações após filtragem: 60;
- Valor de cobertura: 0.722;

Neste teste o número de anotações após filtragem aumenta com o uso de mais duas ontologias e por sua vez aumenta também o valor de cobertura.

Os próximos casos a serem analisados serão os "eventos", testados com análise de termos "texto" e anotados com 1 ontologia e 3 ontologias.

Para o encontro 26999, do paciente 2, anotado com uma ontologia os resultados foram os seguintes:

- Número de termos: 55;
- Número de anotações antes da filtragem: 118;
- Número de anotações após filtragem: 68;
- Valor de cobertura: 0.69;

Os números foram muito semelhantes, houve a diminuição do número de termos, pois o agrupamento "evento" é um subconjunto do agrupamento "paciente". A ontologia utilizada para anotar também foi a NCIT.

No teste feito ao encontro 8779, do paciente 17, anotado com uma ontologia os resultados foram os seguintes:

- Número de termos: 42;
- Número de anotações antes da filtragem: 97;
- Número de anotações após filtragem: 55;
- Valor de cobertura: 0.717;

Neste teste o número de termos baixou, passou de 43 no agrupamento "paciente" do paciente 17, para 42 no agrupamento "encontro". O valor de cobertura contínua semelhante ao que foi obtido nos testes anteriores.

No caso em que o teste foi feito com 3 ontologias e os dados analisados como "texto", os resultados obtidos para o evento 26999, foram:

- Número de termos: 55;
- Número de anotações antes da filtragem: 221;
- Número de anotações após filtragem: 72;
- Valor de cobertura: 0.701;

Mais uma vez houve um aumento do valor de cobertura e a diferença deste valor para o valor médio de cobertura do Recommender é notória, o Recommender teve uma média de 0.903, nos dados guardados na base de dados, e neste teste o valor de cobertura foi de 0.701.

Os dados obtidos para o encontro 8779, do paciente 17, analisado com "texto" e anotado com 3 ontologias foram os seguintes:

- Número de termos: 42;
- Número de anotações antes da filtragem: 179;
- Número de anotações após filtragem: 59;
- Valor de cobertura: 0.727;

Os próximos testes a serem analisados foram realizados da mesma forma que os anteriores mas a única diferença está na forma como os termos foram analisados, foram analisados como "palavras-chave".

O primeiro teste a ser analisado será o agrupamento "paciente" do paciente 2, em que os termos foram analisados como "palavras-chave" e anotados com a NCIT, os resultados foram os seguintes:

- Número de termos: 57;
- Número de anotações antes da filtragem: 13;
- Número de anotações após filtragem: 13;
- Valor de cobertura: 0.236;

Os resultados do teste foram muito diferentes quando comparados com os resultados do teste para agrupamento "paciente" e analise "texto" para o mesmo paciente e o mesmo número de ontologias. O número de anotações foi inferior e o valor de cobertura também inferior. Segundo o Recommender, o valor de cobertura médio para esta situação seria de 0.739, o que é muito superior ao obtido, a diferença deste dois valores deve-se à forma como as anotações são feitas e como o valor de cobertura é calculado pelo Recommender.

Para o paciente 17, os resultados foram os seguintes:

- Número de termos: 43;

- Número de anotações antes da filtragem: 9;
- Número de anotações após filtragem: 9;
- Valor de cobertura: 0.209;

O número de anotações e valor de coberturas obtidos neste teste quando comparados com os resultados do teste simétrico, são inferiores. O valor de cobertura médio calculado pelo Recommender também é superior ao valor de cobertura para este caso.

O próximo teste a ser analisado será o paciente 2, onde os termos foram agrupados por paciente, analisados como "palavras-chave" e anotados com 3 ontologias:

- Número de termos: 57;
- Número de anotações antes da filtragem: 21;
- Número de anotações após filtragem: 13;
- Valor de cobertura: 0.228;

Como é possível observar com o aumento do número de ontologias o valor de cobertura não aumentou, pois as novas ontologias não adicionaram anotações com conceitos diferentes daqueles que a primeira ontologia não tivesse já anotado, contudo o número de anotações aumentou.

O próximo teste a ser analisado é o teste feito ao agrupamento "paciente" do paciente 17, onde os termos foram analisados como "palavras-chave" e anotados com 3 ontologias, os resultados foram os seguintes:

- Número de termos: 43;
- Número de anotações antes da filtragem: 16;
- Número de anotações após filtragem: 9;
- Valor de cobertura: 0.209;

Estes resultados são iguais ao do teste anterior. E para ambos os casos, o valor de cobertura médio calculado pelo Recommender, que é de 0.94, é superior ao valor de cobertura obtido no EHRannotator para ambos os casos.

Os próximos testes a serem analisados são os "eventos" dos pacientes anteriores. O primeiro evento a ser analisado é o 26999 do paciente 2, os dados foram analisados como "palavras-chave" e anotados com uma ontologia.

- Número de termos: 55;
- Número de anotações antes da filtragem: 13;
- Número de anotações após filtragem: 13;
- Valor de cobertura: 0.236;

Para o evento 8779, do paciente 17, os resultados foram os seguintes:

- Número de termos: 42;
- Número de anotações antes da filtragem: 9;
- Número de anotações após filtragem: 9;
- Valor de cobertura: 0.214;

Os resultados dos mesmos testes só que anotados com 3 ontologias foram os seguintes:

Evento do paciente 2:

- Número de termos: 55;
- Número de anotações antes da filtragem: 21;
- Número de anotações após filtragem: 13;
- Valor de cobertura: 0.236;

Evento do paciente 17:

- Número de termos: 42;
- Número de anotações antes da filtragem: 16;
- Número de anotações após filtragem: 9;
- Valor de cobertura: 0.214;